DOCUMENT RESUME

ED 071 533

FL 003 905

TITLE	Status Report on Speech Research, No. 27, July-September 1971.
INSTITUTION	Haskins Labs., New Haven, Conn.
SPONS AGENCY	National Inst. of Child Health and Human Development (NIH), Bethesda, Md.; National Inst. of Dental Research (NIH), Bethesda, Md.; Office of Naval Research, Washington, D.C. Information Systems Research.
REPORT NO	SR-27-1971
PUB DATE	Oct 71
NOTE	211p.
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	Acoustic Phonetics; Articulation (Speech); Artificial Speech; *Communication (Thought Transfer); Distinctive Features; Error Patterns; Information Processing; Language Development; Language Patterns; *Language Research; *Language Skills; Listening; Memory; Physiology; *Reading; Research Methodology; Spectrograms; *Speech; Stimuli; Written Language

ABSTRACT

This report contains fourteen papers on a wide range of current topics and experiments in speech research, ranging from the relationship between speech and reading to questions of memory and perception of speech sounds. The following papers are included: "How Is Language Conveyed by Speech?;" "Reading, the Linguistic Process, and Linguistic Awareness;" "Misreading: A Search for Causes;" "Language codes and Memory Codes;" "Speech Cues and Sign Stimuli;" "On the Evolution of Human Language;" "Distinctive Features and Laryngeal Control;" "Auditory and Linguistic Processes in the Perception of Intonation Contours;" "Glottal Modes in Consonant Distinctions;" "Voice Timing in Korean Stops;" "Interactions between Linguistic and Nonlinguistic Processing;" "Perception of Linguistic and Nonlinguistic Dimensions of Dichotic Stimuli;" "Dichotic Backward Masking of Complex Sounds;" and "On the Nature of Categorical Perception of Speech Sounds." A list of recent publications, reports, oral papers, and theses is included. (VM)

FILMED FROM BEST AVAILABLE COPY

ED 071533

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

SR-27 (1971)

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 July - 30 September 1971

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use.)

FL003 905

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Haskins Laboratories, Inc. 270 Crown Street New Haven, Conn. 06510	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP N/A

3. REPORT TITLE

Status Report on Speech Research, No. 27, July-September 1971

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Interim Scientific Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Staff of Haskins Laboratories; Franklin S. Cooper, P.I.

6. REPORT DATE October 1971	7a. TOTAL NO. OF PAGES 211	7b. NO. OF REFS 364

8a. CONTRACT OR GRANT NO. ONR Contract N00014-67-A-0129-0001 b. NIDR: Grant DE-01774 NICHD: Grant HD-01994 c. NIH/DRFR: Grant FR-5596 VA/PSAS Contract V-1005M-1253 d. NICHD Contract NIH-71-2420	9a. ORIGINATOR'S REPORT NUMBER(S) SR-27 (1971) 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* None

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.*

11. SUPPLEMENTARY NOTES N/A	12. SPONSORING MILITARY ACTIVITY See No. 8

13. ABSTRACT

This report (for 1 July-30 September) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts and extended reports cover the following topics:

- -How is Language Conveyed by Speech?
- -Reading, the Linguistic Process, and Linguistic Awareness
- -Misreading: A Search for Causes
- -Language Codes and Memory Codes
- -Speech Cues and Sign Stimuli
- -On the Evolution of Human Language
- -Distinctive Features and Laryngeal Control
- -Auditory and Linguistic Processes in the Perception of Intonation Contours
- -Glottal Modes in Consonant Distinctions
- -Voice Timing in Korean Stops
- -Interactions Between Linguistic and Nonlinguistic Processing
- -Perception of Linguistic and Nonlinguistic Dimensions of Dichotic Stimuli
- -Dichotic Backward Masking of Complex Sounds
- -On the Nature of Categorical Perception of Speech Sounds

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED
Security Classification

3/D PPSO 13152

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech production Speech perception Speech synthesis Speech cues Reading Language codes Evolution of language Intonation Laryngeal function (in language) Dichotic listening Maskins, auditory Coding, linguistic						

DD FORM 1 NOV 65 1473 (BACK)
S/N 0101-807-6821

UNCLASSIFIED
Security Classification

A-31409

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

Information Systems Branch, Office of Naval Research
Contract N00014-67-A-0129-0001
Req. No. NR 048-225

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration
Contract V-1005M-1253

National Institutes of Health
General Research Support Grant FR-5596

National Institute of Child Health and Human Development
Contract NIH-71-2420

CONTENTS

I. Manuscripts and Extended Reports

Introductory Note. 1

How is Language Conveyed by Speech? -- Franklin S. Cooper 3

Reading, the Linguistic Process, and Linguistic Awareness -- Ignatius G. Mattingly. 23

Misreading: A Search for Causes -- Donald Shankweiler and Isabelle Y. Liberman 35

Language Codes and Memory Codes -- Alvin M. Liberman, Ignatius G. Mattingly, and Michael T. Turvey 59

Speech Cues and Sign Stimuli -- Ignatius G. Mattingly 89

On the Evolution of Human Language -- Philip Lieberman 113

Distinctive Features and Laryngeal Control -- Leigh Lisker and Arthur S. Abramson 133

Auditory and Linguistic Processes in the Perception of Intonation Contours -- Michael Studdert-Kennedy and Kerstin Hadding 153

Glottal Modes in Consonant Distinctions -- Leigh Lisker and Arthur S. Abramson 175

Voice Timing in Korean Stops -- Arthur S. Abramson and Leigh Lisker 179

Interactions Between Linguistic and Nonlinguistic Processing -- Ruth S. Day and Charles C. Wood. 185

Perception of Linguistic and Nonlinguistic Dimensions of Dichotic Stimuli -- Ruth S. Day, James E. Cutting, and Paul M. Copeland 193

Dichotic Backward Masking of Complex Sounds -- C. J. Darwin 199

ABSTRACT: On the Nature of Categorical Perception of Speech Sounds -- David Bob Pisoni 209

ERRATUM: Letter Confusions and Reversals of Sequence in the Beginning Reader -- Isabelle Y. Liberman, Donald Shankweiler, Charles Orlando, Katherine S. Harris, and Fredericka B. Berti 211

II. Publications and Reports 213

INTRODUCTORY NOTE TO STATUS REPORT 27

The first three papers in this Status Report were presented at an invitational conference sponsored by NICHD on the Relationships between Speech and Learning to Read, A.M. Liberman and J.J. Jenkins were the co-chairmen of the conference, which was held at Belmont, Elkridge, Maryland May 16-19, 1971. The conference was divided into three sessions dealing with three closely related topics: (1) the relationship between the terminal signals--written characters or speech sounds--and the linguistic information they convey; (2) the actual processing of information in the linguistic signals and the multiple recodings of these signals; (3) the developmental aspects of reading and speech perception.

The three papers reproduced here with the kind permission of the publisher were presented by staff members of Haskins Laboratories. "How is Language Conveyed by Speech?" by F.S. Cooper was presented at the first session; "Reading, the Linguistic Process, and Linguistic Awareness," by I.G. Mattingly, at the second session; and "Misreading: A Search for Causes," by D.P. Shankweiler and I.Y. Liberman, at the third session. These papers, together with other papers given at the Conference and an Introduction by the co-chairmen, will appear in a book edited by J.F. Kavanagh and I.G. Mattingly. The book, tentatively entitled Language by Ear and by Eye: The Relationships between Speech and Reading, will be published by M.I.T. Press.

How is Language Conveyed by Speech?*

Franklin S. Cooper
Haskins Laboratories, New Haven

In a conference on the relationships between speech and learning to read, it is surely appropriate to start with reviews of what we now know about speech and writing as separate modes of communication. Hence the question now before us: How is language conveyed by speech? The next two papers will ask similar questions about writing systems, both alphabetic and nonalphabetic. The similarities and differences implied by these questions need to be considered not only at performance levels, where speaking and listening are in obvious contrast with writing and reading, but also at the competence levels of spoken and written language. Here, the differences are less obvious, yet they may be important for reading and its successful attainment by the young child.

In attempting a brief account of speech as the vehicle for spoken language, it may be useful first to give the general point of view from which speech and language are here being considered. It is essentially a process approach, motivated by the desire to use experimental findings about speech to better understand the nature of language. So viewed, language is a communicative process of a special--and especially remarkable--kind. Clearly, the total process of communicating information from one person to another involves at least the three main operations of production, transmission, and reception. Collectively, these processes have some remarkable properties: open-endedness, efficiency, speed, and richness of expression. Other characteristics that are descriptive of language processes *per se*, at least when transmission is by speech, include the existence of semantically "empty" elements and a hierarchical organization built upon them; furthermore, as we shall see, the progression from level to level involves restructuring operations of such complexity that they truly qualify as encodings rather than encipherings. The encoded nature of the speech signal is a topic to which we shall give particular attention since it may well be central to the relationship between speech and learning to read.

The Encoded Nature of Speech

It is not intuitively obvious that speech really is an encoded signal or, indeed, that it has special properties. Perhaps speech seems so simple because it is so common: everyone uses it and had done so since early childhood. In fact, the universality of spoken language and its casual acquisition by the young child--even the dullard--are among its most remarkable, and least understood, properties. They set it sharply apart from written language: reading and writing are far from universal, they are acquired only later by formal instruction, and even special instruction often proves ineffective with an otherwise normal child. Especially revealing are the problems of children who lack one of the sensory capacities--vision or hearing--for dealing with language. One finds that blindness is no bar to the effective use of spoken language, whereas deafness severely impedes the mastery of written language, though vision is still intact. Here is further and dramatic evidence that spoken language has a special status not shared by written language. Perhaps, like walking, it comes naturally, whereas skiing does not but can be learned. The nature of the underlying differences between spoken and written language, as well as of the similarities, must surely be relevant to our concern with learning to read. Let us note then that spoken language and written language differ, in addition to the obvious ways, in their relationship to the human being--in the degree to which they may be innate, or at least compatible with his mental machinery.

---

* Paper presented at the Conference on Communicating by Language--The Relationships between Speech and Learning to Read, at Belmont, Elkridge, Maryland, 16-19 May 1971. To appear in *Language by Ear and by Eye: The Relationships between Speech and Reading*, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

Is this compatibility evident in other ways, perhaps in special properties of the speech signal itself? Acoustically, speech is complex and would not qualify by engineering criteria as a clean, definitive signal. Nevertheless, we find that human beings can understand it at rates (measured in bits per second) that are five to ten times as great as for the best engineered sounds. We know that this is so from fifty years of experience in trying to build machines that will read for the blind by converting letter shapes to distinctive sound shapes (Coffey, 1963; Cooper, 1950; Studdert-Kennedy and Cooper, 1966); we know it also--and we know that practice is not the explanation--from the even longer history of telegraphy. Likewise, for speech production, we might have guessed from everyday office experience that speech uses special tricks to go so fast. Thus, even slow dictation will leave an expert typist far behind; the secretary, too, must resort to tricks such as shorthand if she is to keep pace.

Comparisons of listening and speaking with reading and writing are more difficult, though surely relevant to our present concern with what is learned when one learns to read. We know that, just as listening can outstrip speaking, so reading can go faster than writing. The limit on listening to speech appears to be about 400 words per minute (Orr et al., 1965), though it is not yet clear whether this is a human limit on reception (or comprehension) or a machine limit beyond which the process used for time compression has seriously distorted the speech signal. Limits on reading speed are even harder to determine and to interpret, in part because reading lends itself to scanning as listening does not. Then, too, reading has its star performers who can go several times as fast as most of us. But, aside from these exceptional cases, the good reader and the average listener have limiting rates that are roughly comparable. Is the reader, too, using a trick? Perhaps the same trick in reading as in listening?

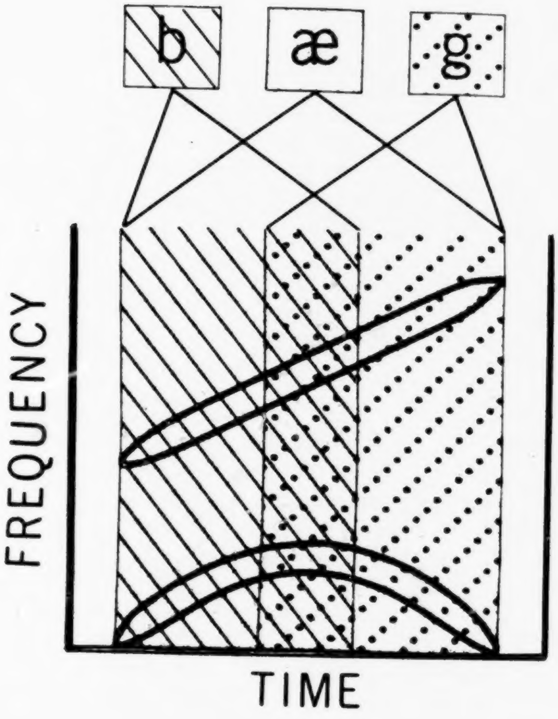
For speech, we are beginning to understand how the trick is done. The answers are not complete, nor have they come easily. But language has proved to be vulnerable to experimental attack at the level of speech, and the insights gained there are useful guides in probing higher and less accessible processes. Much of the intensive research on speech that was sparked by the emergence of sound spectrograms just after World War II was, in a sense, seduced by the apparent simplicities of acoustic analysis and phonemic representation. The goal seemed obvious: it was to find acoustic invariants in speech that matched the phonemes in the message. Although much was learned about the acoustic events of speech, and which of them were essential cues for speech perception, the supposed invariants remained elusive, just as did such promised marvels as the phonetic typewriter. The reason is obvious, now that it is understood: the speech signal was assumed to be an acoustic cipher, whereas it is, in fact, a code.

The distinction is important here as it is in cryptography from which the terms are borrowed: "cipher" implies a one-to-one correspondence between the minimal units of the original and final messages; thus, in Poe's story, "The Goldbug," the individual symbols of the mysterious message stood for the separate letters of the instructions for finding the treasure. In like manner, speech was supposed--erroneously--to comprise a succession of acoustic invariants that stood for the phonemes of the spoken message. The term "code" implies a different and more complex relationship between original and final message. The one-to-one relationship between minimal units has disappeared, since it is the essence of encoding that the original message is restructured (and usually shortened) in ways that are prescribed by an encoding algorithm or mechanism. In commercial codes, for example, the "words" of the final message may all be six-letter groups, regardless of what they stand for. Corresponding units of the original message might be a long corporate name, a commonly used phrase, or a single word or symbol. The restructuring, in this case, is done by substitution, using a code book. There are other methods of encoding--more nearly like speech--which restructure the message in a more or less continuous manner, hence, with less variability in the size of unit on which the encoder operates. It may then be possible to find rough correspondences between input and output elements, although the latter will be quite variable and dependent on context. Further, a shortening of the message may be achieved by collapsing it so that there is temporal overlap of the original units; this constitutes parallel transmission in the sense that there is, at every instant of time, information in the output about several units of the input. A property of such codes is that the output is no longer segmentable, i.e., it cannot be divided into pieces that match units of the input. In this sense also the one-to-one relationship has been lost in the encoding process.

The restructuring of spoken language has been described at length by Liberman et al. (1967). An illustration of the encoded nature of the speech can be seen in Figure 1, from a recent article (Liberman, 1970). It shows a schematic spectrogram that will, if turned back into sound by a speech synthesizer, say "bag" quite clearly. This is a simpler display of frequency, time, and intensity than one would find in a spectrogram of the word as spoken by a human being, but it captures the essential pattern. The figure shows that the influence of the initial and final consonants extend so far into the vowel that they overlap even with each other, and that the vowel influence extends throughout the syllable. The meaning of "influence" becomes clear when one examines comparable patterns for syllables with other consonants or another vowel: thus, the pattern for "gag" has a U-shaped second formant, higher at its center than the midpoint of the second formant shown for "bag"; likewise changing the vowel, as in "bog," lowers the frequency of the second formant not only at the middle of the syllable but at the beginning and end as well.

Clearly, the speech represented by these spectrographic patterns is not an acoustic cipher, i.e., the physical signal is not a succession of sounds

Parallel Transmission of Phonetic Segments
After Encoding (by the Rules of Speech)
to the Level of Sound



(From Liberman, 1970, p. 309.)

Fig. 1

that stand for phonemes. There is no place to cut the syllable "bag" that will isolate separate portions for "b" and "a" and "g." The syllable is carrying information about all of them at the same time (parallel transmission), and each is affected by its neighbors (context dependence). In short, the phonetic string has been restructured, or encoded, into a new element at the acoustic level of the speech signal.

But is speech the only part of language that is encoded? Liberman's article, from which the illustration was drawn, asserts that comparable processes operate throughout language; that the encoding of speech and the transformations of syntactic and phonological structures are broadly similar and equally a part of the grammar. Thus, Figure 2 from the same article shows diagramatically the kind of restructuring and temporal compression that occurs in the syntactic conversion between deep and surface structure. Conventional orthography is used to represent the three deep-structure sentences and the single composite sentence at the surface. Again, there are overlapping domains, and compactness has been bought at the price of substantial changes in structure.

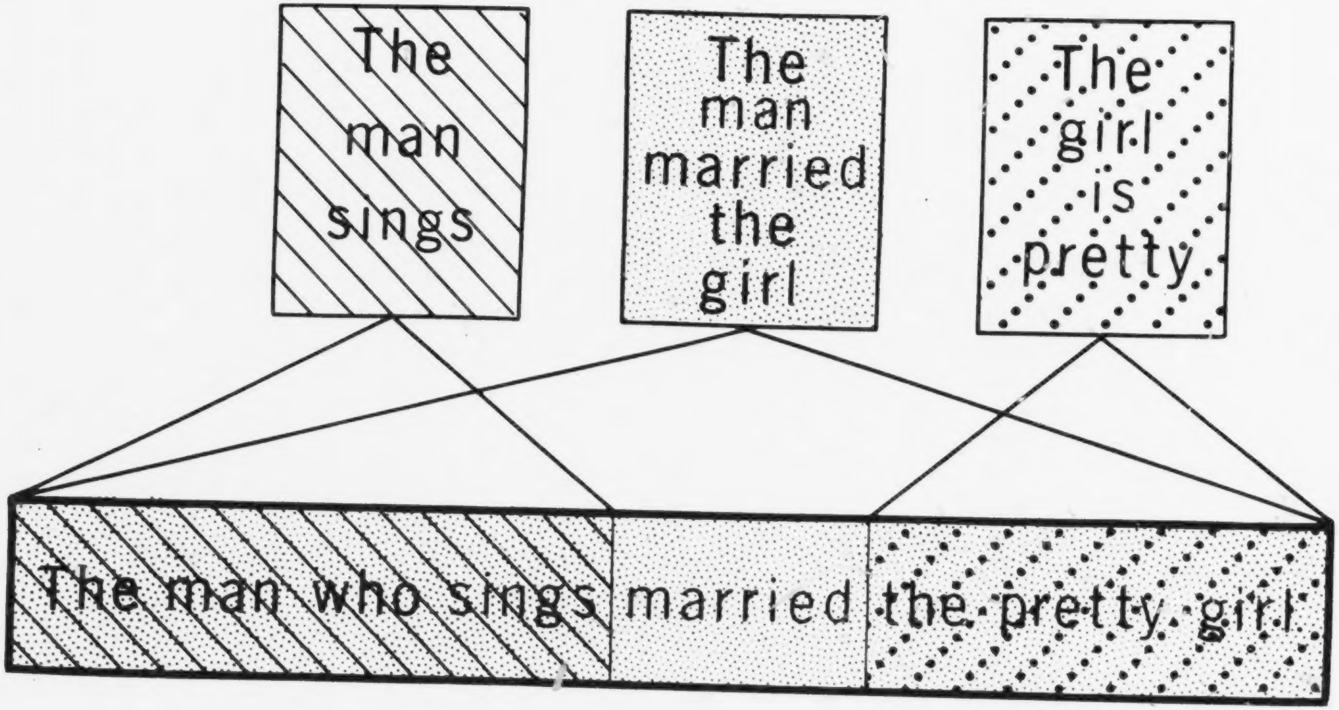
Encoding and Decoding

We see then, in all of spoken language, a very substantial degree of encoding. Why should this be so? Does it serve a purpose, or is it merely an unavoidable consequence of man's biological nature, or both? We have seen, in speech, that there is a temporal telescoping of the phonetic string into syllables and that this speeds communication; also, at the level of syntax, that there is a comparable collapsing of the deep structures into surface structures, with further gains in speed. Moreover, there are cognitive advantages that may be even more important, and that may explain why the encoding seems to have been done in stages, resulting in an hierarchical structure for language. George Miller (1956) has given us an account of how the magic of encoding lets us deal with substantial quantities of information in spite of limited memory capacity.

These are impressive advantages, but the price seems very high. We would suppose, from the foregoing, that the task of the person who listens to speech is staggeringly difficult: he must somehow deal with a signal that is an encoding of an encoding of an encoding.... Indeed, the difficulties are very real, as many people have discovered in trying to build speech recognizers or automatic parsing programs. But the human being does it so easily that we can only suppose he has access to full knowledge (even if implicit) of the coding relationships. These relationships, or a model of the processes by which the encoding is done, could fully rationalize for him the involved relation of speech signal to underlying message and so provide the working basis for his personal speech decoder (liberman, 1970).

Our primary interest is, of course, in how speech is perceived, since this is where we would expect to find relationships with reading and its acquisition. It is not obvious that a person's implicit knowledge of how his own speech is produced might help to explain how another's speech can be perceived. Actually, we think that it does, although, even without such a premise, one would need to know how the encoding is done, since that is what the decoder must undo. So, before we turn to a discussion of how speech is perceived, let us first consider how it is produced.

Parallel Transmission of Deep-Structure Segments
After Encoding (by the Rules of Syntax)
to the Level of Surface Structure



(From Liberman, 1970, p. 310.)

Fig. 2

The Making of Spoken Language

Our aim is to trace in a general way the events that befall a message from its inception as an idea to its expression as speech. Much will be tentative, or even wrong, at the start but can be more definite in the final stages of speech production. There, where our interest is keenest, the experimental evidence is well handled by the kinds of models often used by communications engineers. This, together with the view that speech is an integral part of language, suggests that we might find it useful to extrapolate a communications model to all stages of language production.

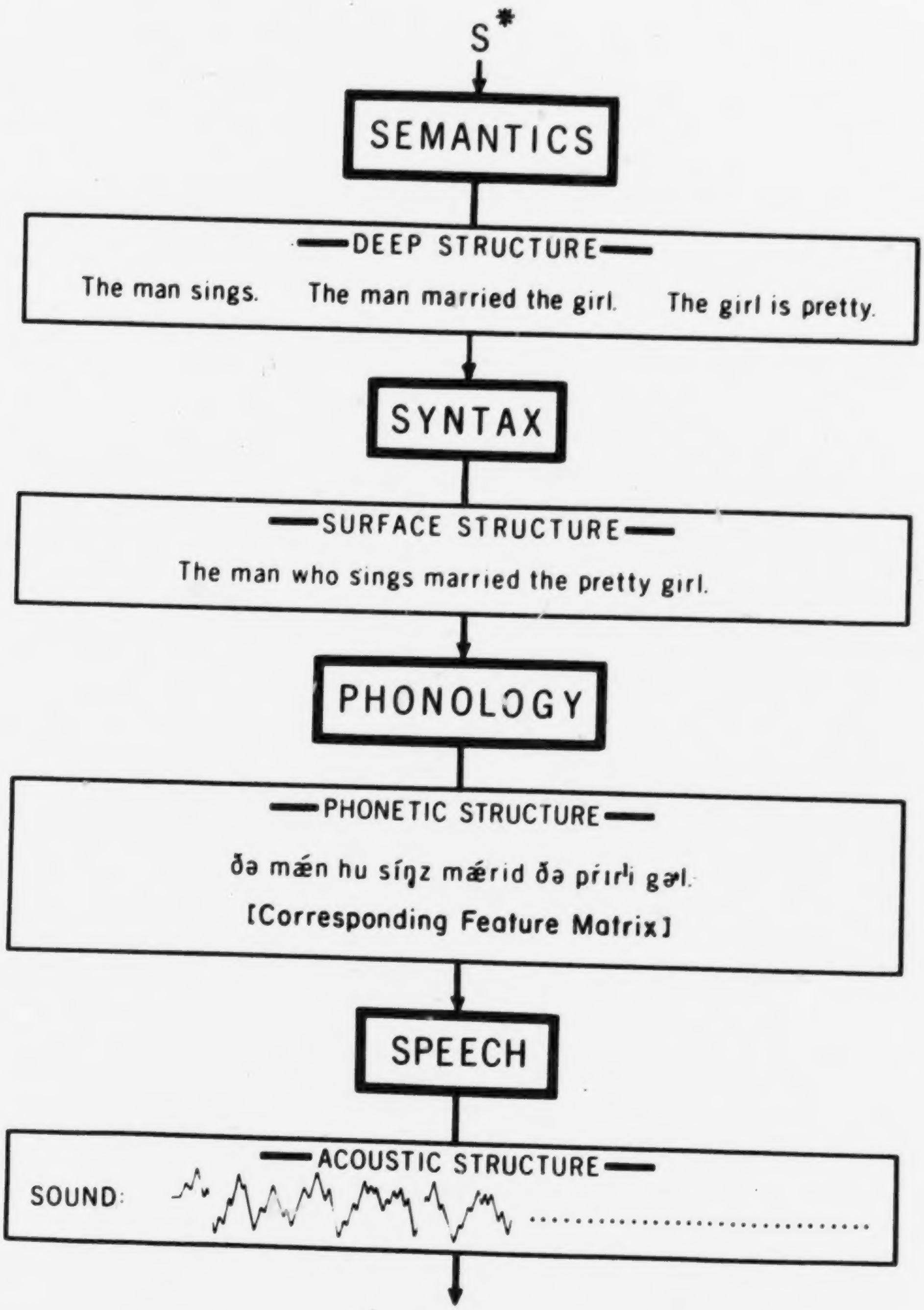
The conventional block diagram in Figure 3 can serve as a way of indicating that a message (carried on the connecting lines) undergoes sequential transformations as it travels through a succession of processors. The figure shows a simple, linear arrangement of the principal processors (the blocks with heavy outlines) that are needed to produce spoken language and gives descriptions (in the blocks with light outlines) of the changing form of the message as it moves from processor to processor on its way to the outside world. The diagram is adapted from Liberman (1970) and is based (in its central portions) on the general view of language structure proposed by Chomsky and his colleagues (Chomsky, 1957, 1965; Chomsky and Miller, 1963). We can guess that a simple, linear process of this kind will serve only as a first approximation; in particular, it lacks the feedback and feedforward paths that we would expect to find in a real-life process.

We know quite well how to represent the final (acoustic) form of a message--assumed, for convenience, to be a sentence--but not how to describe its initial form. S*, then, symbolizes both the nascent sentence and our ignorance about its prelinguistic form. The operation of the semantic processor is likewise uncertain, but its output should provide the deep structure--corresponding to the three simple sentences shown for illustration--on which syntactic operations will later be performed. Presumably, then, the semantic processor will somehow select and rearrange both lexical and relational information that is implicit in S*, perhaps in the form of semantic feature matrices.

The intermediate and end results of the next two operations, labeled Syntax and Phonology, have been much discussed by generative grammarians. For present purposes, it is enough to note that the first of them, syntactic processing, is usually viewed as a two-stage operation, yielding firstly a phrase-structure representation in which related items have been grouped and labeled, and secondly a surface-structure representation which has been shaped by various transformations into an encoded string of the kind indicated in the figure (again, by its plain English counterpart). Some consequences of the restructuring of the message by the syntactic processor are that (1) a linear sequence has been constructed from the unordered cluster of units in the deep structure and (2) there has been the telescoping of the structure, hence encoding, that we saw in Figure 2 and discussed in the previous section.

Further restructuring of the message occurs in the phonological processor. It converts (encodes) the more or less abstract units of its input into a time-ordered array of feature states, i.e., a matrix showing the state of each feature for each phonetic event in its turn. An alternate representation would

A Process Model for the Production of Spoken Language



The intended message flows down through a series of processors (the blocks with heavy outlines). Descriptions are given (in the blocks with light outlines) of the changing form of the message as it moves from processor to processor. (Adapted from Liberman, 1970, p. 305.)

Fig. 3

be a phonetic string that is capable of emerging at least into the external world as a written phonetic transcription.

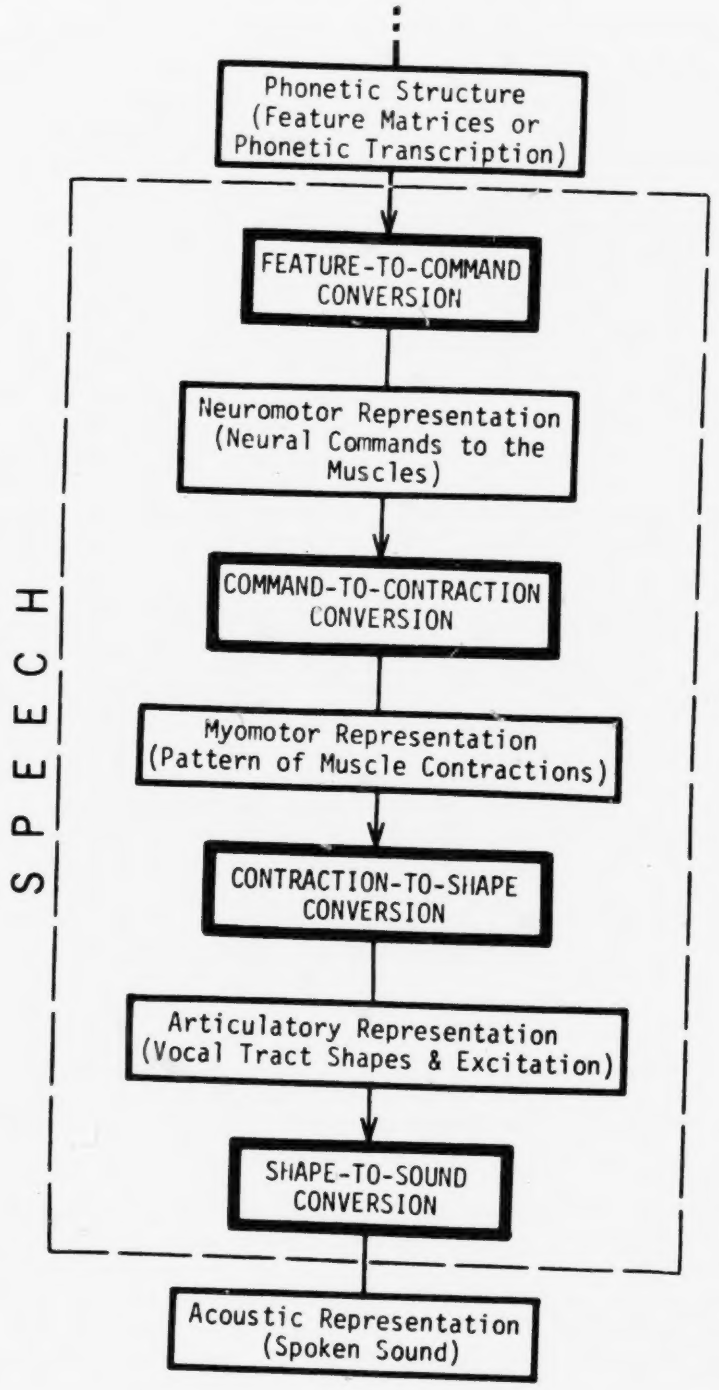
This is about where contemporary grammar stops, on the basis that the conversion into speech from either the internal or external phonetic representation--although it requires human intervention--is straightforward and essentially trivial. But we have seen, with "bag" of Figure 1 as an example, that the spoken form of a message is a heavily encoded version of its phonetic form. This implies processing that is far from trivial--just how far is suggested by Figure 4, which shows the major conversions required to transform an internal phonetic representation into the external acoustic waveforms of speech. We see that the speech processor, represented by a single block in Figure 3, comprises several subprocessors, each with its own function: firstly, the abstract feature matrices of the phonetic structure must be given physiological substance as neural signals (commands) if they are to guide and control the production of speech; these neural commands then bring about a pattern of muscle contractions; these, in turn, cause the articulators to move and the vocal tract to assume a succession of shapes; finally, the vocal-tract shape (and the acoustic excitation due to air flow through the glottis or other constrictions) determines the spoken sound.

Where, in this sequence of operations, does the encoding occur? If we trace the message upstream--processor by processor, starting from the acoustic outflow--we find that the relationships between speech waveform and vocal-tract shape are essentially one-to-one at every moment and can be computed, though the computations are complex (Fant, 1960; Flanagan, 1965). However, at the next higher stop--the conversion of muscle contractions into vocal-tract shapes--there is substantial encoding: each new set of contractions starts from whatever configuration and state of motion already exist as the result of preceding contractions, and it typically occurs before the last set is ended, with the result that the shape and motion of the tract at any instant represent the merged effects of past and present events. This alone could account for the kind of encoding we saw in Figure 1, but whether it accounts for all of it, or only a part, remains to be seen.

We would not expect much encoding in the next higher conversion--from neural command to muscle contraction--at least in terms of the identities of the muscles and the temporal order of their activation. However, the contractions may be variable in amount due to preplanning at the next higher level or to local adjustment, via gamma-efferent feedback, to produce only so much contraction as is needed to achieve a target length.

At the next higher conversion--from features to neural commands--we encounter two disparate problems: one involves functional, physiological relationships very much like the ones we have just been considering, except that their location in the nervous system puts them well beyond the reach of present experimental methods. The other problem has to do with the boundary between two kinds of description. A characteristic of this boundary is that the feature matrix (or the phonetic transcription) provided by the phonological processor is still quite abstract as compared with the physiological type of feature that is needed as an input to the feature-to-command conversion. The simple case--and perhaps the correct one--would be that the two sets of features are fully congruent, i.e., that the features at the output of the phonology will

Internal Structure of the Speech Processor



Again, the message flows from top to bottom through successive processors (the blocks with heavy outlines), with intermediate descriptions given (in the blocks with light outlines).

Fig. 4

map directly onto the distinctive components of the articulatory gestures. Failing some such simple relationship, translation or restructuring would be required in greater or lesser degree to arrive at a set of features which are "real" in a physiological sense. The requirement is for features rather than segmental (phonetic) units, since the output of the conversion we are considering is a set of neural commands that go in parallel to the muscles of several essentially independent articulators. Indeed, it is only because the features--and the articulators--operate in this parallel manner that speech can be fast even though the articulators are slow.

The simplistic hypothesis noted above, i.e., that there may be a direct relationship between the phonological features and characteristic parts of the gesture, has the obvious advantage that it would avoid a substantial amount of encoding in the total feature-to-command conversion. Even so, two complications would remain. In actual articulation, the gestures must be coordinated into a smoothly flowing pattern of motion which will need the cooperative activity of various muscles (in addition to those principally involved) in ways that depend on the current state of the gesture, i.e., in ways that are context dependent. Thus, the total neuromotor representation will show some degree of restructuring even on a moment-to-moment basis. There is a further and more important sense in which encoding is to be expected: if speech is to flow smoothly, a substantial amount of preplanning must occur, in addition to moment-by-moment coordination. We know, indeed, that this happens for the segmental components over units at least as large as the syllable and for the suprasegmentals over units at least as large as the phrase. Most of these coordinations will not be marked in the phonetic structure and so must be supplied by the feature-to-command conversion. What we see at this level, then, is true encoding over a longer span of the utterance than the span affected by lower-level conversions and perhaps some further restructuring even within the shorter span.

There is ample evidence of encoding over still longer stretches than those affected by the speech processor. The sentence of Figure 2 provides an example--one which implies processor and conversion operations that lie higher in the hierarchical structure of language than does speech. There is no reason to deny these processors the kind of neural machinery that was assumed for the feature-to-command conversion; however, we have very little experimental access to the mechanisms at these levels, and we can only infer the structure and operation from behavioral studies and from observations of normal speech.

In the foregoing account of speech production, the emphasis has been on processes and on models for the various conversions. The same account could also be labeled a grammar in the sense that it specifies relationships between representations of the message at successive stages. It will be important, in the conference discussions on the relationship of speaking to reading, that we bear in mind the difference between the kind of description used thus far--a process grammar--and the descriptions given, for example, by a generative transformational grammar. In the latter case, one is dealing with formal rules that relate successive representations of the message, but there is now no basis for assuming that these rules mirror actual processes. Indeed, proponents of generative grammar are careful to point out that such an implication is not intended; unfortunately, their terminology is rich in words that seem to imply active operations and cause-and-effect relationships. This can lead to confusion in discussions about the processes that are involved in listening and reading and how they make contact with each other. Hence, we shall need to use the descriptions of rule-based grammars with some care in dealing with experimental data and model mechanisms that reflect, however crudely, the real-life processes of language behavior.

Perception of Speech

We come back to an earlier point, slightly rephrased: how can perceptual mechanisms possibly cope with speech signals that are as fast and complex as the production process has made them? The central theme of most current efforts to answer that question is that perception somehow borrows the machinery of production. The explanations differ in various ways, but the similarities substantially outweigh the differences.

There was a time, though, when acoustic processing *per se* was thought to account for speech perception. It was tempting to suppose that the patterns seen in spectrograms could be recognized *as patterns* in audition just as in vision (Cooper et al., 1951). On a more analytic level, the distinctive features described by Jakobson, Fant, and Halle (1963) seemed to offer a basis for direct auditory analysis, leading to recovery of the phoneme string. Also at the analytic level, spectrographic patterns were used extensively in a search for the acoustic cues for speech perception (Liberman, 1957; Liberman et al. 1967; Stevens and House, in press). All of these approaches reflected, in one way or another, the early faith we have already mentioned in the existence of acoustic invariants in speech and in their usefulness for speech recognition by man or machine.

Experimental work on speech did not support this faith. Although the search for the acoustic cues was successful, the cues that were found could be more easily described in articulatory than in acoustic terms. Even "the locus," as a derived invariant, had a simple articulatory correlate (Delattre et al., 1955). Although the choice of articulation over acoustic pattern as a basis for speech perception was not easy to justify since there was almost always a one-to-one correspondence between the two, there were occasional exceptions to this concurrence which pointed to an articulatory basis, and these were used to support a motor theory of speech perception. Older theories of this kind had invoked actual motor activity (though perhaps minimal in amount) in tracking incoming speech, followed by feedback of sensory information from the periphery to let the listener know what both he and the speaker were articulating. The revised formulation that Liberman (1957) gave of a motor theory to account for the data about acoustic cues was quite general, but it explicity excluded any reference to the periphery as a necessary element:

> All of this [information about exceptional cases] strongly suggests...that speech is perceived by reference to articulation--that is, that the articulatory movements and their sensory effects mediate between the acoustic stimulus and the event we call perception. In its extreme and old-fashioned form, this view says that we overtly mimic the incoming speech sounds and then respond to the appropriate receptive and tactile stimuli that are produced

> by our own articulatory movements. For a variety of reasons such an extreme position is wholly untenable, and if we are to deal with perception in the adult, we must assume that the process is somehow short-circuited--that is, that the reference to articulatory movements and their sensory consequences must somehow occur in the brain without getting out into the periphery. (p. 122)

A further hypothesis about how the mediation might be accomplished (Liberman et al., 1968) supposes that there is a spread of neural activity within and among sensory and motor networks so that some of the same interlocking nets are active whether one is speaking (and listening to his own speech) or merely listening to speech from someone else. Hence, the neural activity initiated by listening, as it spreads to the motor networks, could cause the whole process of production to be started up just as it would be in speaking (but with spoken output suppressed); further, there would be the appropriate interaction with those same neural mechanisms--whatever they are --by which one is ordinarily aware of what he is saying when he himself is the speaker. This is equivalent, insofar as awareness of another's speech is concerned, to running the production machinery backward, assuming that the interaction between sensory and motor networks lies at about the linguistic level of the features (represented neurally, of course) but that the linkage to awareness is at some higher level and in less primitive terms. Whether or not such an hypothesis about the role of neural mechanisms in speaking and listening can survive does not really affect the main point of a more general motor theory, but it can serve here as an example of the kind of machinery that is implied by a motor theory and as a basis for comparison with the mechanisms that serve other theoretical formulations.

The model for speech perception proposed by Stevens and Halle (1967; Halle and Stevens, 1962) also depends heavily on mechanisms of production. The analysis-by-synthesis procedure was formulated initially in computer terms, though funtional parallels with biological mechanisms were also considered. The computer-like description makes it easier to be specific about the kinds of mechanisms that are proposed but somewhat harder to project the model into a human skull.

It is unnecessary to trace in detail the operation of the analysis-by-synthesis model, but Figure 5, from Stevens's (1960) paper on the subject, can serve as a reminder of much that is already familiar. The processing within the first loop (inside the dashed box) compares spectral information received from the speech input and held in a temporary store with spectral information generated by a model of the articulatory mechanism (Model I). This model receives its instructions from a control unit that generates articulatory states and uses heuristic processes to select a likely one on the basis of past history and the degree of mismatch that is reported to it by a comparator. The articulatory description that is used by Model I (and passed on to the next loop) might have any one of several representations: acoustical, in terms of the normal modes of vibration of the vocal tract; or anatomical, descriptive of actual vocal-tract configurations; or neurophysiological, specifying control signals that would cause the vocal tract to change shape. Most of Stevens's discussion deals with vocal-tract configuration (and excitation); hence, he treats comparisons in the second loop as between input configurations (from the preceding loop) and those generated

Analysis-by-Synthesis Model of Speech Recognition

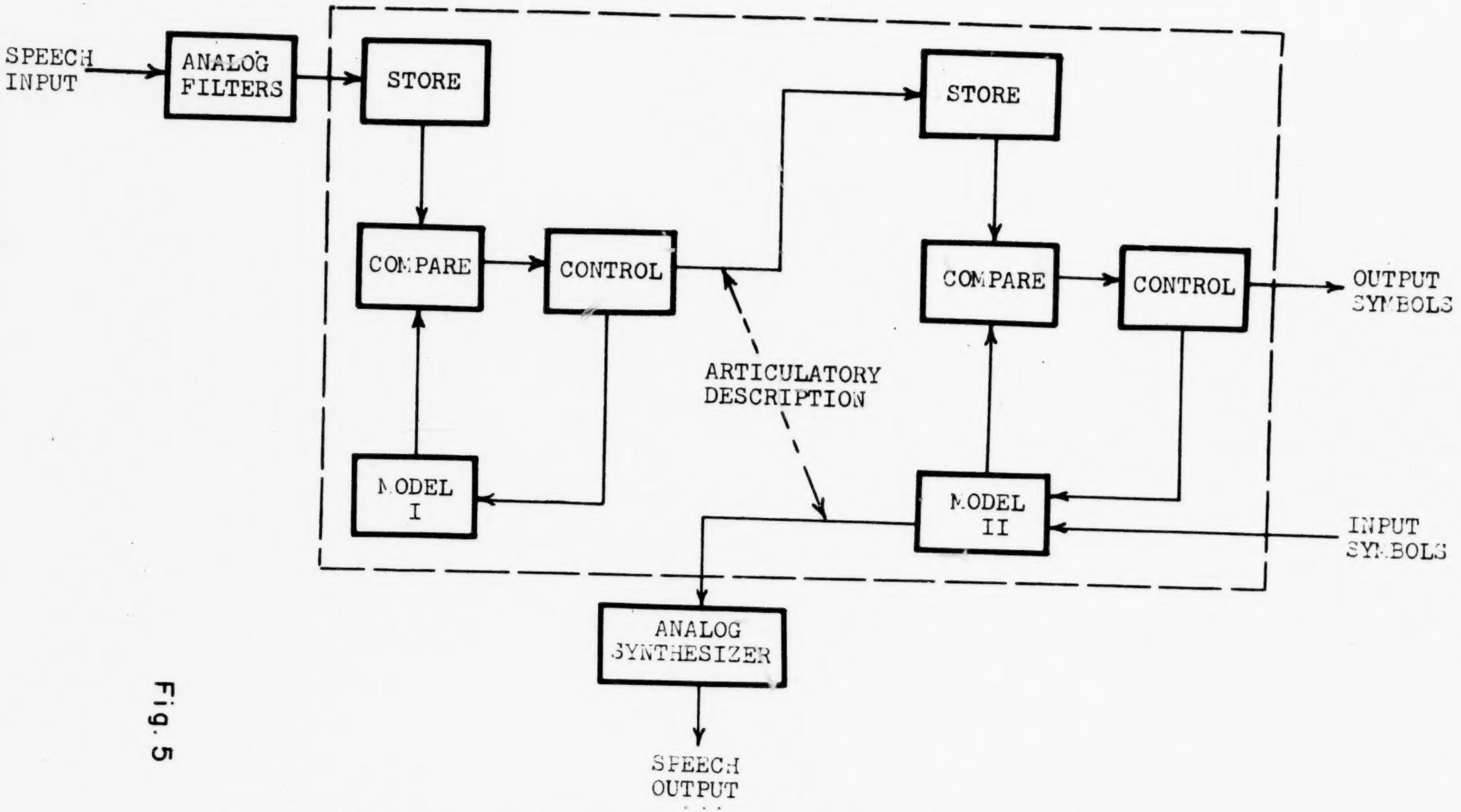


Fig. 5

The acoustic signal enters at the upper left and is "recognized" in the form of a string of phonetic symbols that leave at center right. Model I stores the rules that relate articulatory descriptions to speech spectra, and Model II stores the rules that relate phonetic symbols to articulatory descriptions. Model II can serve also to generate a speech output from an input of phonetic symbols. (From Stevens, 1960, p. 52.)

by an articulatory control (Model II) that could also be used to drive a vocal-tract-analog synthesizer external to the analysis-by-synthesis system. There is a second controller, again with dual functions: it generates a string of phonetic elements that serve as the input to Model II, and it applies heuristics to select, from among the possible phonetic strings, one that will maintain an articulatory match at the comparator.

A virtue of the analysis-by-synthesis model is that its components have explicit functions, even though some of these component units are bound to be rather complicated devices. The comparator, explicit here, is implicit in a neural network model in the sense that some neural nets will be aroused --and others will not--on the basis of degree of similarity between the firing patterns of the selected nets and the incoming pattern of neural excitation. Comparisons and decisions of this kind may control the spread of excitation throughout all levels of the neural mechanism, just as a sophisticated guessing game is used by the analysis-by-synthesis model to work its way, stage by stage, to a phonetic representation--and presumably on upstream to consciousness. In short, the two models differ substantially in the kinds of machinery they invoke and the degree of explicitness that this allows in setting forth the underlying philosophy: they differ very little in the reliance they put on the mechanisms of production to do most of the work of perception.

The general point of view of analysis-by-synthesis is incorporated in the constructionist view of cognitive processes in general, with speech perception as an interesting special case. Thus, Neisser, in the introduction to Cognitive Psychology, says

> The central assertion is that seeing, hearing, and remembering are all acts of construction, which may make more or less use of stimulus information depending on circumstances. The constructive processes are assumed to have two stages, of which the first is fast, crude, wholistic, and parallel while the second is deliberate, attentive, detailed, and sequential. (1967, p. 10).

It seems difficult to come to grips with the specific mechanisms (and their functions) that the constructivists would use in dealing with spoken language to make the total perceptual process operate. A significant feature, though, is the assumption of a two-stage process, with the constructive act initiated on the basis of rather crude information. In this, it differs from both of the models that we have thus far considered. Either model could, if need be, tolerate input data that are somewhat rough and noisy, but both are designed to work best with "clean" data, since they operate first on the detailed structure of the input and then proceed stepwise toward a more global form of the message.

Stevens and House (in press) have proposed a model for speech perception that is, however, much closer to the constructionist view of the process than was the early analysis-by-synthesis model of Figure 5. It assumes that spoken language has evolved in such a way as to use auditory distinctions and attributes that are well matched to optimal performances of the speech generating mechanism; also, that the adult listener has command of a catalog of correspondences between the auditory attributes and the articulatory gestures (of approximately syllabic length) that give rise to them when he is a speaker. Hence, the listener can, by consulting his catalog, infer the speaker's gestures. However, some further analysis is needed to arrive at the phonological features, although their correspondence with articulatory events will often be quite close. In any case, this further analysis allows the "construction" (by a control unit) of a tentative hypothesis about the sequence of linguistic units and the constituent structure of the utterance. The hypothesis, plus the generative rules possessed by every speaker of the language, can then yield an articulatory version of the utterance. In perception, actual articulation is suppressed but the information about it goes to a comparator where it is matched against the articulation inferred from the incoming speech. If both versions match, the hypothesized utterance is confirmed; if not, the resulting error signal guides the control unit in modifying the hypothesis. Clearly, this model employs analysis-by-synthesis principles. It differs from earlier models mainly in the degree of autonomy that the control unit has in constructing hypotheses and in the linguistic level and length of utterance that are involved.

The approach to speech perception taken by Chomsky and Halle (1968) also invokes analysis by synthesis, with even more autonomy in the construction of hypotheses; thus,

> We might suppose...that a correct description of perceptual processes would be something like this. The hearer makes use of certain cues and certain expectations to determine the syntactic structure and semantic content of an utterance. Given a hypothesis as to its syntactic structure--in particular its surface structure--he uses the phonological principles that he controls to determine a phonetic shape. The hypothesis will then be accepted if it is not too radically at variance with the acoustic material, where the range of permitted discrepancy may vary widely with conditions and many individual factors. Given acceptance of such a hypothesis, what the hearer "hears" is what is internally generated by the rules. That is, he will "hear" the phonetic shape determined by the postulated syntactic structure and the internalized rules. (p. 24)

This carries the idea of analysis by synthesis in constructionist form almost to the point of saying that only the grosser cues and expectations are needed for perfect reception of the message (as the listener would have said it), unless there is a gross mismatch with the input information, which is otherwise largely ignored. This extension is made explicit with respect to the perception of stress. Mechanisms are not provided, but they would not be expected in a rule-oriented account.

In all the above approaches, the complexities inherent in the acoustic signal are dealt with indirectly rather than by postulating a second mechanism (at least as complex as the production machinery) to perform a straightforward auditory analysis of the spoken message. Nevertheless, some analysis is needed to provide neural signals from the auditory system for use in generating hypotheses and in error comparisons at an appropriate stage of the production process. Obviously, the need for analysis will be least if the comparisons are made as far down in the production process as possible. It may be, though, that direct auditory analysis plays a larger role. Stevens (1971) has postulated that the analysis is done (by auditory property detectors) in terms of acoustic features that qualify as distinctive features of the language, since they are both inherently distinctive and directly related to stable articulatory states. Such an auditory analysis might not yield complete information about the phonological features of running speech, but enough, nevertheless, to activate analysis-by-synthesis operations. Comparisons could then guide the listener to self-generation of the correct message. Perhaps Dr. Stevens will give us an expanded account of this view of speech perception in his discussion of the present paper.

All these models for perception, despite their differences, have in common a listener who actively participates in producing speech as well as in listening to it in order that he may compare his internal utterances with the incoming one. It may be that the comparators are the functional component of central interest in using any of these models to understand how reading is done by adults and how it is learned by children. The level (or levels) at which comparisons are made--hence, the size and kind of unit compared--determines how far the analysis of auditory (and visual) information has to be carried, what must be held in short-term memory, and what units of the child's spoken language he is aware of--or can be taught to be aware of--in relating them to visual entities.

Can we guess what these units might be, or at least what upper and lower bounds would be consistent with the above models of the speech process? It is the production side of the total process to which attention would turn most naturally, given the primacy ascribed to it in all that has been said thus far. We have noted that the final representation of the message, before it leaves the central nervous system on its way to the muscles, is an array of features and a corresponding (or derived) pattern of neural commands to the articulators. Thus, the features would appear to be the smallest units of production that are readily available for comparison with units derived from auditory analysis. But we noted also that smoothly flowing articulation requires a restructuring of groups of features into syllable- or word-size units, hence, these might serve instead as the units for comparison. In either case, the lower bound on duration would approximate that of a syllable.

The upper bound may well be set by auditory rather than productive processes. Not only would more sophisticated auditory analysis be required to match higher levels--and longer strings--of the message as represented in production, but also the demands on short-term memory capacity would increase. The latter alone could be decisive, since the information rate that is needed to specify the acoustic signal is very high--indeed, so high that some kind of auditory processing must be done to allow the storage of even word-length stretches. Thus, we would guess that the capacity of short-term memory for purely auditory forms of the speech signal would set an upper bound on duration hardly greater than that of words or short phrases. The limits, after conversion to linguistic form, are however substantially longer, as they would have to be for effective communication.

Intuitively, these minimal units seem about right: words, syllables, or short phrases seem to be what we say, and hear ourselves saying, when we talk. Moreover, awareness of these as minimal units is consistent with the reference-to-production models we have been considering, since all of production that lies below the first comparator has been turned over to bone-and-muscle mechanisms (aided, perhaps, by gamma-efferent feedback) and so is inaccessible in any direct way to the neural mechanisms responsible for awareness. As adults, we know how to "analyze" speech into still smaller (phonetic) segments, but this is an acquired skill and not one to be expected of the young child.

Can it be that the child's level of awareness of minimal units in speech is part of his problem in learning to read? Words should pose no serious problem so long as the total inventory remains small and the visual symbols are sufficiently dissimilar. But phonic methods, to help him deal with a larger vocabulary, may be assuming an awareness that he does not have of the phonetic segments of speech, especially his own speech. If so, perhaps learning to read comes second to learning to speak and listen with awareness. This is a view that Dr. Mattingly will, I believe, develop in depth. It can serve here as an example of the potential utility of models of the speech process in providing insights into relationships between speech and learning to read.

In Conclusion

The emphasis here has been on the processes of speaking and listening as integral parts of the total process of communicating by spoken language. This concentration on speech reflects both its role as a counterpart to reading and its accessibility via experimentation. The latter point has not been exploited in the present account, but it is nonetheless important as a reason for focusing on this aspect of language. Most of the unit processors that were attributed to speech in the models we have been discussing can, indeed, be probed experimentally: thus, with respect to the production of speech, electromyography and cinefluorography have much to say about how the articulators are moved into the observed configurations, and sound spectrograms give highly detailed accounts of the dynamics of articulation and acoustic excitation; examples with respect to speech perception include the use of synthetic speech in discovering the acoustic cues inherent in speech, and of dichotic methods for evading peripheral effects in order to overload the central processor and so to study its operation. Several of the papers to follow will deal with comparable methods for studying visual information processing. Perhaps the emphasis given here to processes and to the interdependence of perception and production will provide a useful basis for considering the linkages between reading and speech.

REFERENCES

Chomsky, N. (1957) Syntactic Structures. (The Hague: Mouton).

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: M.I.T. Press).

Chomsky, N. and Halle, M. (1968) The Sound Pattern of English. (New York: Harper and Row).

Chomsky, N. and Miller, G.A. (1963) Introduction to the formal analysis of natural languages. In Handbook of Mathematical Psychology, R.D. Luce, R.R. Bush, and E. Galanter, eds. (New York: Wiley).

Coffey, J.L. (1963) The development and evaluation of the Battelle Aural Reading Device. In Proceedings of the International Congress on Technology and Blindness. (New York: American Foundation for the Blind).

Cooper, F.S. (1950) Research on reading machines for the blind. In Blindness: Modern Approaches to the Unseen Environment, P.A. Zahl, ed. (Princeton, N.J.: Princeton University Press).

Cooper, F.S., Liberman, A.M., and Borst, J.M. (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc. Nat. Acad. Sci. 37, 318-28.

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955) Acoustic loci and transitional cues for consonants. J. acoust. Soc. Am. 27, 769-773.

Fant, C.G.M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Flanagan, J.L. (1965) Speech Analysis, Synthesis and Perception. (New York: Academic Press).

Halle, M. and Stevens, K.N. (1962) Speech recognition: A model and a program for research. IRE Trans. Info. Theory IT-8, 155-59. [Also in The Structure of Language, J.A. Fodor and J.J. Katz, eds. (Englewood Cliffs, N.J.: Prentice-Hall, 1964).]

Jakobson, R., Fant, C.G.M., and Halle, M. (1963) Preliminaries to Speech Analysis. (Cambridge, Mass.: M.I.T. Press).

Liberman, A.M. (1957) Some results of research on speech perception. J. acoust. Soc. Am. 29, 117-123.

Liberman, A.M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A.M., Cooper, F.S., Studdert-Kennedy, M., Harris, K.S., and Shankweiler, D.P. (1968) On the efficiency of speech sounds. Zeits. f. Phonetik, Sprachwissenschaft u. Kommunikationsforschung 21, 21-32.

Miller, G.A. (1956) The magical number seven, plus or minus two, or, some limits on our capacity for processing information. Psychol. Rev. 63, 81-96.

Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).

Orr, D.B., Friedman, H.L. and Williams, J.C.C. (1965) Trainability of listening comprehension of speeded discourse. J. educ. Psychol. 56, 148-156.

Stevens, K.N. (1960) Toward a model for speech recognition. J. acoust. Soc. Am. 32, 47-55.

Stevens, K.N. (1971) Perception of phonetic segments: Evidence from phonology, acoustics and psychoacoustics. In Perception of Language, D.L. Horton and J.J. Jenkins, eds. (Columbus, Ohio: Chas. Merrill).

Stevens, K.N., and Halle, M. (1967) Remarks on analysis by synthesis and distinctive features. In Models for the Perception of Speech and Visual Form, W. Wathen-Dunn, ed. (Cambridge, Mass.: M.I.T. Press).

Stevens, K.N. and House, A.S. (in press) Speech perception. In Foundations of Modern Auditory Theory, Vol. 2, J. Tobias, ed. (New York: Academic Press).

Studdert-Kennedy, M. and Cooper, F.S. (1966) High-performance reading machines for the blind: Psychological problems, technological problems, and status. In Proceedings of the International Conference on Sensory Devices for the Blind, R. Dufton, ed. (London: St. Dunstan's).

Reading, the Linguistic Process, and Linguistic Awareness[*]

Ignatius G. Mattingly[+]
Haskins Laboratories, New Haven

Reading, I think, is a rather remarkable phenomenon. The more we learn about speech and language, the more it appears that linguistic behavior is highly specific. The possible forms of natural language are very restricted; its acquisition and function are biologically determined (Chomsky, 1965). There is good reason to believe that special neural machinery is intricately linked to the vocal tract and the ear, the output and input devices used by all normal human beings for linguistic communication (Liberman et al., 1967). It is therefore rather surprising to find that a minority of human beings can also perform linguistic functions by means of the hand and the eye. If we had never observed actual reading or writing we would probably not believe these activities to be possible. Faced with the fact, we ought to suspect that some special kind of trick is involved. What I want to discuss is this trick, and what lies behind it--the relationship of the process of reading a language to the processes of speaking and listening to it. My view is that this relationship is much more devious than it is generally assumed to be. Speaking and listening are primary linguistic activities, reading is a secondary and rather special sort of activity which relies critically upon the reader's awareness of these primary activities.

The usual view, however, is that reading and listening are parallel processes. Written text is input by eye, and speech, by ear, but at as early a stage as possible, consistent with this difference in modality, the two inputs have a common internal representation. From this stage onward, the two processes are identical. Reading is ordinarily learned later than speech; this learning is therefore essentially an intermodal transfer, the attainment of skill in doing visually what one already knows how to do auditorily. As Fries (1962:xv) puts it

> Learning to read...is not a process of learning new or other language signals than those the child has already learned. The language signals are all the same. The difference lies in the medium through which the physical stimuli make contact with his nervous system. In "talk" the physical stimuli of the language signals make their contact by means of sound waves received by the ear. In reading, the physical stimuli of the same language signals consist of graphic shapes that make their contact with the nervous system through light waves received by the eye. The process of learning to read is the process of transfer from the auditory signs for language signals which the child has already learned, to the new visual signs for the same signals.

---

[*] Paper presented at the Conference on Communicating by Language--The Relationships between Speech and Learning to Read, at Belmont, Elkridge, Maryland, 16-19 May 1971. To appear in Language by Ear and by Eye: The Relationships between Speech and Reading, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

[+] Also University of Connecticut, Storrs.

Something like this view appears to be shared by many who differ about other aspects of reading, even about the nature of the linguistic activity involved. Thus Bloomfield (1942), Fries, and others assume that the production and perception of speech are inversely related processes of encoding and decoding, and take the same view of writing and reading. They believe that the listener extracts the phonemes or "unit speech sounds" from speech, forms them into morphemes and sentences, and decodes the message. Similarly, the reader produces, in response to the text, either audible unit speech sounds or, in silent reading, "internal substitute movements" (Bloomfield, 1942: 103) which he treats as phonemes and so decodes the message. Fries's model is similar to Bloomfield's except that his notion of a phoneme is rather more abstract; it is a member of a set of contrasting elements, conceptually distinct from the medium which conveys it. This medium is the acoustic signal for the listener, the line of print for the reader. For Fries as for Bloomfield, acquisition of both the spoken and written language requires development of "high-speed recognition responses" to stimuli which "sink below the threshold of attention" (Fries, 1962:xvi) when the responses have become habitual.

More recently, however, the perception of speech has come to be regarded by many as an "active" process basically similar to speech production. The listener understands what is said through a process of "analysis by synthesis" (Stevens and Halle, 1967). Parallel proposals have accordingly been made for reading. Thus Hochberg and Brooks (1970) suggest that once the reader can visually discriminate letters and letter groups and has mastered the phoneme-grapheme correspondences of his writing system, he uses the same hypothesis-testing procedure in reading as he does in listening [Goodman's (1970) view of reading as a "psycholinguistic guessing game" is a similar proposal]. Though the model of linguistic processing is different from that of Bloomfield and Fries, the assumption of a simple parallel between reading and listening remains, and the only differences mentioned are those assignable to modality, for example, the use which the reader makes of peripheral vision, which has no analog in listening.

While it is clear that reading somehow employs the same linguistic processes as listening, it does not follow that the two activities are directly analogous. There are, in fact, certain differences between the two processes which cannot be attributed simply to the difference of modality and which therefore make difficulties for the notion of a straightforward intermodal parallel. Most of these differences have been pointed out before, notably by Liberman et al. (1967) and Liberman (in Kavanagh, 1968). But I think reconsideration of them will help us to arrive at a better understanding of reading.

To begin with, listening appears to be a more natural way of perceiving language than reading; "listening is easy and reading is hard" (Liberman, in Kavanagh, 1968:119). We know that all living languages are spoken languages and that every normal child gains the ability to understand his native speech as part of a maturational process of language acquisition. In fact we must suppose that, as a prerequisite for language acquisition, the child has some kind of innate capability to perceive speech. In order to extract from the utterances of others the "primary linguistic data" which he needs for acquisition, he must have a "technique for representing input signals" (Chomsky, 1965: 30).

In contrast, relatively few languages are written languages. In general, children must be deliberately taught to read and write, and despite this teaching, many of them fail to learn. Someone who has been unable to acquire language by listening--a congenitally deaf child, for instance--will hardly be able to acquire it through reading; on the contrary, as Liberman and Furth (in Kavanagh, 1968) point out, a child with a language deficit owing to deafness will have great difficulty learning to read properly.

The apparent naturalness of listening does not mean that it is in all respects a more efficient process. Though many people find reading difficult, there are a few readers who are very proficient: in fact, they read at rates well over 2,000 words per minute with complete comprehension. Listening is always a slower process: even when speech is artificially speeded up in a way which preserves frequency relationships, 400 words per minute is about the maximum possible rate (Orr et al., 1965). It has often been suggested (e.g., Bever and Bower, 1966) that high-speed readers are somehow able to go directly to a deep level of language, omitting the intermediate stages of processing to which other readers and all listeners must presumably have recourse.

Moreover, the form in which information is presented is basically different in reading and in listening. The listener is processing a complex acoustic signal in which the speech cues that constitute significant linguistic data are buried. Before he can use these cues, the listener has to "demodulate" the signal: that is, he has to separate the cues from the irrelevant detail. The complexity of this task is indicated by the fact that no scheme for speech recognition by machine has yet been devised which can perform it properly. The demodulation is largely unconscious; as a rule, a listener is unable to perceive the actual acoustic form of the event which serves as a cue unless it is artificially excised from its speech context (Mattingly et al., 1971). The cues are not discrete events, well separated in time or frequency; they blend into one another. We cannot, for instance, realistically identify a certain instant as the ending of a formant transition for an initial consonant and the beginning of the steady state of the following vowel.

The reader, on the other hand, is processing a series of symbols which are quite simply related to the physical medium which conveys them. The task of demodulation is straightforward: the marks in black ink are information; the white paper is background. The reader has no particular difficulty in seeing the letters as visual shapes if he wants to. In printed text, the symbols are discrete units. In cursive writing, of course, one can slur together the symbols to a surprising degree without loss of legibility. But though they are deformed, the cursive symbols remain essentially discrete. It makes sense to view cursive writing as a string of separate symbols connected together for practical convenience; it makes no sense at all to view the speech signal in this way.

That these differences in form are important is indicated by the difficulty of reading a visual display of the speech signal, such as a sound spectrogram, or of listening to text coded in an acoustic alphabet, e.g., Morse code or any of the various acoustic alphabets designed to aid the blind (Studdert-Kennedy and Liberman, 1963; Coffey, 1963). We know that a spectrogram contains most of the essential linguistic information, for it can be converted back to acoustic form without much loss of intelligibility (Cooper, 1950). Yet reading a spectrogram is very slow work at best, and at worst, impossible. Similarly, text coded in an acoustic alphabet contains the same information as print, but a listener can follow it only if it is presented at a rate which is very slow compared to a normal speaking rate.

These facts are certainly not quite what we should predict if reading and listening were simply similar processes in different modalities. The relative advantage of the eye with alphabetic text, to be sure, may be attributed to its apparent superiority over the ear as a data channel; but then why should the eye do so poorly with visible speech? We can only infer that some part of the neural speech processing machinery must be accessible through the ear but not through the eye.

There is also a difference in the linguistic content of the information available to the listener and the reader. The speech cues carry information about the phonetic level of language, the articulatory gestures which the speaker must have made--or more precisely, the motor commands which lead to those gestures (Lisker et al., 1962). Written text corresponds to a different level of language. Chomsky (1970) makes the important observation that conventional orthography, that of English in particular, is, roughly speaking, a morphophonemic transcription; in the framework of generative grammar, it corresponds fairly closely to a surface-structure phonological representation. I think this generalization can probably be extended to include all practical writing systems, despite their apparent variety. The phonological level is quite distinct from the phonetic level, though the two are linked in each language by a system of phonological rules. The parallel between listening and reading was plausible in part because of the failure of structural linguistics to treat these two linguistic levels as the significant ones: both speech perception and reading were taken to be phonemic. Chomsky (1964) and Halle (1959), however, have argued rather convincingly that the phonemic level of the structuralists has no proper linguistic significance, its supposed functions being performed at either the phonological or the phonetic level.

Halwes (in Kavanagh, 1968:160) has observed:

> It seems like a good bet that since you have all this apparatus in the head for understanding language that if you wanted to teach somebody to read, you would arrange a way to get the written material input to the system that you have already got for processing spoken language and at as low a level as you could arrange to do that, then let the processing of the written material be done by the mechanisms that are already in there.

I think that Halwes's inference is a reasonable one, and since the written text does not, in fact, correspond to the lowest possible level, the problem is with his premise, that reading and listening are simply analogous processes.

There is, furthermore, a difference in the way the linguistic content and the information which represents it are related. As Liberman (in Kavanagh, 1968: 120) observes, "speech is a complex code, print a simple cipher." The nature of the speech code by which the listener deduces articulatory behavior from acoustic events is determined by the characteristics of the vocal tract. The code is complex because the physiology and acoustics of the vocal tract are complex. It is also a highly redundant code: there are, typically, many acoustic cues for a single bit of phonetic information. It is, finally, a universal code, because all human vocal tracts have similar properties. By comparison, writing is, in principle, a fairly simple mapping of units of the phonological representation--morphemes or phonemes or syllables--into written symbols. The complications which do occur are not determined by the nature of what is being represented: they are historical accidents. By comparison with the speech code, writing is a very economical mapping; typically, many bits of phonological information are carried by a single symbol. Nor is there any inherent relationship between the form of written symbols and the corresponding phonological units; to quote Liberman once more (in Kavanagh, 1968:121), "only one set of sounds will work, but there are many equally good alphabets."

The differences we have listed indicate that even though reading and listening are both clearly linguistic and have an obvious similarity of function, they are not really parallel processes. I would like to suggest a rather different interpretation of the relationship of reading to language. This interpretation depends on a distinction between primary linguistic activity itself and the speaker-hearer's awareness of this activity.

Following Miller and Chomsky (1963), Stevens and Halle (1967), Neisser (1967), and others, I view primary linguistic activity, both speaking and listening, as essentially creative or synthetic. When a speaker-hearer "synthesizes" a sentence, the products are a semantic representation and a phonetic representation which are related by the grammatical rules of his language, in the sense that the generation of one entails the generation of the other. The speaker must synthesize and so produce a phonetic representation for a sentence which, according to the rules, will have a particular required semantic representation; the listener, similarly, must synthesize a sentence which matches a particular phonetic representation, in the process recovering its semantic representation. It should be added that synthesis of a sentence does not necessarily involve its utterance. One can think of a sentence without actually speaking it; one can rehearse or recall a sentence.

Since we are concerned with reading and not with primary linguistic activity as such, we will not attempt the difficult task of specifying the actual process of synthesis. We merely assume that the speaker-hearer not only knows the rules of his language but has a set of strategies for linguistic performance. These strategies, relying upon context as well as upon information about the phonetic (or semantic) representation to be matched, are powerful enough to insure that the speaker-hearer synthesizes the "right" sentence most of the time.

Having synthesized some utterance, whether in the course of production or perception, the speaker-hearer is conscious not only of a semantic experience (understanding the utterance) and perhaps an acoustic experience (hearing the speaker's voice) but also of experience with certain intermediate linguistic processes. Not only has he synthesized a particular utterance, he is also aware in some way of having done so and can reflect upon this linguistic experience as he can upon his experiences with the external world.

If language were in great part deliberately and consciously learned behavior, like playing the piano, this would hardly be very surprising. We would suppose that development of such linguistic awareness was needed in order to learn language. But if language is acquired by maturation, linguistic awareness seems quite remarkable when we consider how little introspective awareness we have of the intermediate stages of other forms of maturationally acquired motor and perceptual behavior, for example, walking or seeing.

The speaker-hearer's linguistic awareness is what gives linguistics its special advantage in comparison with other forms of psychological investigation. Taking his informant's awareness of particular utterances as a point of departure, the linguist can construct a description of the informant's intuitive competence in his language which would be unattainable by purely behavioristic methods (Sapir, 1949).

However, linguistic awareness is very far from being evenly distributed over all phases of linguistic activity. Much of the process of synthesis takes place well beyond the range of immediate awareness (Chomsky, 1965) and must be determined inferentially--just how much has become clear only recently, as a result of investigations of deep syntactic structure by generative grammarians and of speech perception by experimental phoneticians. Thus the speaker-hearer's knowledge of the deep structure and transformational history of an utterance is evident chiefly from his awareness of the grammaticality of the utterance or its lack of it; he has no direct awareness at all of many of the most significant acoustic cues, which have been isolated by means of perceptual experiments with synthetic speech.

On the other hand, the speaker-hearer has a much greater awareness of phonetic and phonological events. At the phonetic level, he can often detect deviations, even in the case of features which are not distinctive in his language, and this sort of awareness can be rapidly increased by appropriate ear training.

At the phonological (surface-structure) level, not only distinctions between deviant and acceptable utterances, but also reference to various structural units, becomes possible. Words are perhaps most obvious to the speaker-hearer, and morphemes hardly less so, at least in the case of languages with fairly elaborate inflectional and compounding systems. Syllables, depending upon their structural role in the language, may be more obvious than phonological segments. There is far greater awareness of the structural unit than of the structure itself, so that the speaker-hearer feels that the units are simply concatenated. The syntactic bracketing of the phonological representation is probably least obvious.

In the absence of appropriate psycholinguistic data, any ordering of this sort is, of course, very tentative, and in any case, it would be a mistake to overstate the clarity of the speaker-hearer's linguistic awareness and the consistency with which it corresponds to a particular linguistic level. But it is safe to say that, by virtue of this awareness, he has an internal image of the utterance, and this image probably owes more to the phonological level of representation than to any other level.

There appears to be considerable individual variation in linguistic awareness. Some speaker-hearers are not only very conscious of linguistic patterns but exploit their consciousness with obvious pleasure in verbal play, e.g., punning or verbal work (e.g., linguistic analysis). Others seem never to be aware of much more than words and are surprised when quite obvious linguistic patterns are pointed out to them. This variation contrasts markedly with the relative consistency from person to person with which primary linguistic activity is performed. Synthesis of an utterance is one thing; the awareness of the process of synthesis, quite another.

Linguistic awareness is by no means only a passive phenomenon. The speaker-hearer can use his awareness to control, quite consciously, his linguistic activity. Thus he can ask himself to synthesize a number of words containing a certain morpheme, or a sentence in which the same phonological segment recurs repeatedly.

Without this active aspect of linguistic awareness, moreover, much of what we call thinking would be impossible. The speaker-hearer can consciously represent things by names and complex concepts by verbal formulas. When he tries to think abstractly, manipulating these names and concepts, he relies ultimately upon his ability to recapture the original semantic experience. The only way to do this is to resynthesize the utterance to which the name or formula corresponds.

Moreover, linguistic awareness can become the basis of various language-based skills. Secret languages, such as Pig Latin (Halle, 1964) form one class of examples. In such languages a further constraint, in the form of a rule relating to the phonological representation, is artificially imposed upon production and perception. Having synthesized a sentence in English, an additional mental operation is required to perform the encipherment. To carry out the process at a normal speaking rate, one has not only to know the rule but also to have developed a certain facility in applying it. A second class of examples are the various systems of versification. The versifier is skilled in synthesizing sentences which conform not only to the rules of the language but to an additional set of rules relating to certain phonetic features (Halle, 1970). To listen to verse, one needs at least a passive form of this skill so that one can readily distinguish "correct" from "incorrect" lines without scanning them syllable by syllable.

It seems to me that there is a clear difference between Pig Latin, versification, and other instances of language-based skill, and primary linguistic activity itself. If one were unfamiliar with Pig Latin or with a system of versification, one might fail to understand what the Pig Latinist or the versifier was up to, but one would not suppose either of them to be speaking an unfamiliar language. And even after one does get on to the trick, the sensation of engaging in something beyond primary linguistic activity does not disappear. One continues to be aware of a special demand upon our linguistic awareness.

Our view is that reading is a language-based skill like Pig Latin or versification and not a form of primary linguistic activity analogous to listening. From this viewpoint, let us try to give an account, necessarily much oversimplified, of the process of reading a sentence.

The reader first forms a preliminary, quasi-phonological representation of the sentence based on his visual perception of the written text. The form in which this text presents itself is determined not by the actual linguistic information conveyed by the sentence but by the writer's linguistic awareness of the process of synthesizing the sentence, an awareness which the writer wishes to impart to the reader. The form of the text does not consist, for instance, of a tree-structure diagram or a representation of articulatory gestures, but of discrete units, clearly separable from their visual context. These units, moreover, correspond roughly to elements of the phonological representation (in the generative grammarian's sense), and the correspondence between these units and the phonological elements is quite simple. The only real question is whether the writing system being used is such that the units represent morphemes, or syllables, or phonological segments.

Though the text is in a form which appeals to his linguistic awareness, considerable skill is required of the reader. If he is to proceed through the text at a practical pace, he cannot proceed unit by unit. He must have an extensive vocabulary of sight words and phrases acquired through previous reading experience. Most of the time he identifies long strings of units. When this sight vocabulary does fail him, he must be ready with strategies by means of which he can identify a word which is part of his spoken vocabulary and add it to his sight vocabulary or assign a phonological representation to a word altogether unknown to him. To be able to do this he must be thoroughly familiar with the rules of the writing system: the shapes of the characters and the relationship of characters and combinations of characters to the phonology of his language. Both sight words and writing system are matters of convention and must be more or less deliberately learned. While their use becomes habitual in the skilled reader, they are never inaccessible to awareness in the way that much primary linguistic activity is.

The preliminary representation of the sentence will contain only a part of the information in the linguist's phonological representation. All writing systems omit syntactic, prosodic, and junctural information, and many systems make other omissions: for example, phonological vowels are inadequately represented in English spelling and omitted completely in some forms of Semitic writing. Thus the preliminary representation recovered by the reader from the written text is a partial version of the phonological representation: a string of words which may well be incomplete and are certainly not syntactically related.

The skilled reader, however, does not need complete phonological information and probably does not use all of the limited information available to him. The reason is that the preliminary phonological representation serves only to control the next step of the operation, the actual synthesis of the sentence. By means of the same primary linguistic competence he uses in speaking and listening, the reader endeavors to produce a sentence which will be consistent with its context and with this preliminary representation.

In order to do this, he needs, not complete phonological information, but only enough to exclude all other sentences which would fit the context. As he synthesizes the sentence, the reader derives the appropriate semantic representation and so understands what the writer is trying to say.

Does the reader also form a phonetic representation? Though it might seem needless to do so in silent reading, I think he does. In view of the complex interaction between levels which must take place in primary linguistic activity, it seems unlikely that a reader could omit this step at will. Moreover, as suggested earlier, even though writing systems are essentially phonological, linguistic awareness is in part phonetic. Thus, a sentence which is phonetically bizarre--"The rain in Spain falls mainly in the plain," for example--will be spotted by the reader. And quite often, the reason a written sentence appears to be stylistically offensive is that it would be difficult to speak or listen to.

Having synthesized a sentence which fits the preliminary phonological representation, the reader proceeds to the actual recognition of the written text, that is, he applies the rules of the writing system and verifies, at least in part, the sentence he has synthesized. Thus we can, if we choose, think of the reading process as one analysis-by-synthesis loop inside another, the inner loop corresponding to primary linguistic activity and the outer loop to the additional skilled behavior used in reading. This is a dangerous analogy, however, because the nature of both the analysis and the synthesis is very different in the two processes.

This account of reading ties together many of the differences between reading and listening noted earlier: the differences in the form of the input information, the difference in its linguistic content, and the difference in the relationship of form to content. But we have still to explain the two most interesting differences: the relatively higher speeds which can be attained in reading and the relative difficulty of reading.

How can we explain the very high speeds at which some people read? To say that such readers go directly to a semantic representation, omitting most of the process of linguistic synthesis, is to hypothesize a special type of reader who differs from other readers in the nature of his primary linguistic activity, differs in a way which we have no other grounds for supposing possible. As far as I know, no one has suggested that high-speed readers can listen, rapidly or slowly, in the way they are presumed to read. A more plausible explanation is that linguistic synthesis takes place much faster than has been supposed and that the rapid reader has learned how to take advantage of this. The relevant experiments (summarized by Neisser, 1967) have measured the rate at which rapidly articulated or artificially speeded speech can be comprehended and the rate at which a subject can count silently, that is, the rate of "inner speech." But since temporal relationships in speech can only withstand so much distortion, speeded speech experiments may merely reflect limitations on the rate of input. The counting experiment not only used unrealistic material but assumed that inner speech is an essential concomitant of linguistic synthesis. But suppose that the inner speech which so many readers report, and which figures so prominently in the literature on reading, is simply a kind of auditory imagery, dependent upon linguistic awareness of the sentence already synthesized, reassuring but by no means essential (any more than actual utterance or subvocalization) and rather time-consuming. One could then explain the high-speed reader as someone who builds up the preliminary representation efficiently and synthesizes at a very high speed, just as any other reader or speaker-hearer does. But since he is familiar with the nature of the text, he seldom finds it necessary to verify the output of the process of synthesis and spends no time on inner speech. The high speed at which linguistic synthesis occurs is directly reflected in his reading speed. This explanation is admittedly speculative but has the attraction of treating the primary linguistic behavior of all readers as similar and assigning the difference to behavior peculiar to reading.

Finally, why should reading be, by comparison with listening, so perilous a process? This is not the place to attempt an analysis of the causes of dyslexia, but if our view of reading is correct, there is plenty of reason why things should often go wrong. First, we have suggested that reading depends ultimately on linguistic awareness and that the degree of this awareness varies considerably from person to person. While reading does not make as great a demand upon linguistic awareness as, say, solving British crossword puzzles, there must be a minimum level required, and perhaps not everyone possesses this minimum: not everyone is sufficiently aware of units in the phonological representation or can acquire this awareness by being taught. In the special case of alphabetic writing, it would seem that the price of greater efficiency in learning is a required degree of awareness higher than for logographic and syllabary systems, since as we have seen, phonological segments are less obvious units than morphemes or syllables. Almost any Chinese with ten years to spare can learn to read, but there are relatively few such people. In a society where alphabetic writing is used, we should expect more reading successes, because the learning time is far shorter, but proportionately more failures, too, because of the greater demand upon linguistic awareness.

A further source of reading difficulty is that the written text is a grosser and far less redundant representation than speech: one symbol stands for a lot more information than one speech cue, and the same information is not available elsewhere in the text. Both speaker and listener can perform sloppily and the message will get through: the listener who misinterprets a single speech cue will often be rescued by several others. Even a listener with some perceptual difficulty can muddle along. The reader's tolerance of noisy input is bound to be much lower than the listener's, and a person with difficulty in visual perception so mild as not to interfere with most other tasks may well have serious problems in reading.

These problems are both short- and long-term. Not only does the poor reader risk misreading the current sentence, but there is the possibility that his vocabulary of sight words and phrases will become corrupted by bad data and that the strategies he applies when the sight vocabulary fails will be the wrong strategies. In this situation he will build up the preliminary phonological representation not only inaccurately, which in itself might not be so serious, but too slowly, because he is forced to have recourse to his strategies so much of the time. This is fatal, because a certain minimum rate of input seems to be required for linguistic synthesis. We know, from experience with speech slowed by inclusion of a pause after each word, that even when individual words are completely intelligible, it is hard to put the whole sentence together. If only a reader can maintain the required minimum rate of input, many of his perceptual errors can be smoothed over in synthesis: it is no doubt for this reason that most readers manage as well as they do. But if he goes too slowly, he may well be unable to keep up with his own processes of linguistic synthesis and will be unable to make _any_ sense out of what he reads.

Liberman has remarked that reading is parasitic on language (in Kavanagh, 1968). What I have tried to do here, essentially, is to elaborate upon that notion. Reading is seen not as a parallel activity in the visual mode to speech perception in the auditory mode: there are differences between the two activities which cannot be explained in terms of the difference of modality. They can be explained only if we regard reading as a deliberately acquired, language-based skill, dependent upon the speaker-hearer's awareness of certain aspects of primary linguistic activity. By virtue of this linguistic awareness, written text initiates the synthetic linguistic process common to both reading and speech, enabling the reader to get the writer's message and so to recognize what has been written.

REFERENCES

Bever, T.G. and Bower, T.G. (1966) How to read without listening. Project Literacy Reports No. 6, 13-25.

Bloomfield L.(1942) Linguistics and reading. Elementary English Rev., 125-130 & 183-186.

Chomsky, N. (1964) Current Issues in Linguistic Theory. (The Hague: Mouton).

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: M.I.T. Press).

Chomsky, N. (1970) Phonology and reading. In Basic Studies on Reading, Harry Levin and Joanna Williams, eds. (New York: Basic Books).

Coffey, J.L. (1963) The development and evaluation of the Battelle Aural Reading Device. In Proc. Int. Cong. Technology and Blindness. (New York: American Foundation for the Blind).

Cooper, F.S. (1950) Spectrum analysis. J. acoust. Soc. Amer. 22, 761-762.

Fries, C.C. (1962) Linguistics and Reading. (New York: Holt, Rinehart and Winston).

Goodman, K.S. (1970) Reading: A psycholinguistic guessing game. In Theoretical Models and Processes of Reading, Harry Singer and Robert B. Ruddell, eds. (Newark, Del.: International Reading Association).

Halle, M. (1959) The Sound Pattern of Russian. (The Hague: Mouton).

Halle, M. (1964) On the bases of phonology. In The Structure of Language, J.A. Foder and J.J. Katz, eds. (Englewood Cliffs, N.J.: Prentice-Hall).

Halle, M. (1970) On metre and prosody. In Progress in Linguistics, M. Bierwisch and K. Heidolph, eds. (The Hague: Mouton).

Hochberg, J. and Brooks, V. (1970) Reading as an intentional behavior. In Theoretical Models and Processes of Reading, H. Singer and R.B. Ruddell, eds. (Newark, Del.: International Reading Association).

Kavanagh, J.F., ed. (1968) Communicating by Language: The Reading Process. (Bethesda, Md.: National Institute of Child Health and Human Development).

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Lisker, L., Cooper, F.S. and Liberman A.M. (1962) The uses of experiment in language description. Word 18, 82-106.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K. and Halwes, T. (1971) Discrimination in speech and non-speech modes. Cognitive Psychology 2, 131-157.

Miller, G. and Chomsky, N. (1963) Finitary models of language users. In Handbook of Mathematical Psychology, R.D. Luce, R.R. Bush, and E. Galanter, eds. (New York: Wiley).

Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).

Orr, D.B., Friedman, H.L. and Williams, J.C.C. (1965) Trainability of listening comprehension of speeded discourse. J. educ. Psychol. 56, 148-156.

Sapir, E. (1949) The psychological reality of phonemes. In Selected Writings of Edward Sapir in Language, Culture, and Personality, D.G. Mandelbaum, ed. (Berkeley: University of Calif. Press).

Studdert-Kennedy, M. and Liberman, A.M. (1963) Psychological considerations in the design of auditory displays for reading machines. In Proc. Int. Cong. Technology and Blindness. (New York: American Foundation for the Blind).

Stevens, K.N. and Halle, M. (1967) Remarks on analysis by synthesis and distinctive features. In Models for the Perception of Speech and Visual Form, W. Wathen-Dunn, ed. (Cambridge, Mass: M.I.T. Press).

Misreading: A Search for Causes[*]

Donald Shankweiler[+] and Isabelle Y. Liberman[++]

Because speech is universal and reading is not, we may suppose that the latter is more difficult and less natural. Indeed, we know that a large part of the early education of the school child must be devoted to instruction in reading and that the instruction often fails, even in the most favorable circumstances. Judging from the long history of debate concerning the proper methods of teaching children to read (Mathews, 1966), the problem has always been with us. Nor do we appear to have come closer to a solution: we are still a long way from understanding how children learn to read and what has gone wrong when they fail.

Since the child already speaks and understands his language at the time reading instruction begins, the problem is to discover the major barriers in learning to perceive language by eye. It is clear that the first requirement for reading is that the child be able to segregate the letter segments and identify them with accuracy and speed. Some children undoubtedly do fail to learn to recognize letters and are unable to pass on to succeeding stages of learning to read, but as we shall see, there are strong reasons for believing that the principal barriers for most children are not at the point of visual identification of letter shapes. There is no general agreement, however, about the succeeding stages of learning to read, their time course, and the nature of their special difficulties. In order to understand reading and compare it with speech, we need to look closely at the kinds of difficulties the child has when he starts to read, that is, his misreadings, and ask how these differ from errors in repeating speech perceived by ear. In this way, we may begin to grasp why the link between alphabet and speech is difficult.

In the extensive literature about reading since the 1890's there have been sporadic surges of interest in the examination of oral reading errors

---

[*] Paper presented at the Conference on Communicating by Language--The Relationships between Speech and Learning to Read, at Belmont, Elkridge, Maryland, 16-19 May 1971. To appear in Language by Ear and by Eye: The Relationships between Speech and Reading, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

[+] Haskins Laboratories, New Haven, and University of Connecticut, Storrs.

[++] University of Connecticut, Storrs.

Acknowledgments: This work was supported in part by a grant to the University of Connecticut from the U.S. Office of Education (principal investigator, I.Y. Liberman). Many of the ideas expressed here were contributed by our colleagues at Haskins Laboratories in the course of many discussions. A.M. Liberman and L. Lisker read a draft of the paper and made many valuable comments. It is a please to acknowledge their help.

as a means of studying the process of reading acquisition. The history of this topic has been well summarized by Weber (1968), so need not be repeated here. We ourselves set out in many directions when we began our pursuit of errors and we regard our work as essentially exploratory. If we break new ground, it is not by our interest in error patterns nor even in many of our actual findings, but rather in the questions we are asking about them.

Much of the most recent research on reading errors has examined the child's oral reading of connected text (Goodman, 1965, 1968; Schale, 1966; Weber, 1968; Christenson, 1969; Biemiller, 1970). The major emphasis of these studies is therefore on levels beyond the word, though they are concerned to some extent with errors within words. None of these investigations asks what we believe to be a basic question: whether the major barrier to reading acquisition is indeed in reading connected text or whether it may be instead in dealing with words and their components.

We are, in addition, curious to know whether the difficulties in reading are to be found at a visual stage or at a subsequent linguistic stage of the process. This requires us to consider the special case of reversal errors, in which optical considerations are, on the face of it, primary. Our inquiry into linguistic aspects of reading errors then leads us to ask which constituents of words tend to be misread and whether the same ones tend to be misheard. We examine errors with regard to the position of the constituent segments within the word and the linguistic status of the segments in an attempt to produce a coherent account of the possible causes of the error pattern in reading.

We think all the questions we have outlined can be approached most profitably by studying children who are a little beyond the earliest stages of reading instruction. For this reason, we have avoided the first grade and focused, in most of our work, on children of the second and third grades of the elementary school. Though some of the children at this level are well on their way to becoming fluent in reading, a considerable proportion are still floundering and thus provide a sizeable body of errors for examination.

THE WORD AS THE LOCUS OF DIFFICULTY IN BEGINNING READING

One often encounters the claim that there are many children who can read individual words well yet do not seem able to comprehend connected text (Anderson and Dearborn, 1952; Goodman, 1968). The existence of such children is taken to support the view that methods of instruction which stress spelling-to-sound correspondences and other aspects of decoding are insufficient and may even produce mechanical readers who are expert at decoding but fail to comprehend sentences. It may well be that such children do exist; if so, they merit careful study. Our experience suggests that the problem is rare, and that poor reading of text with little comprehension among beginning readers is usually a consequence of reading words poorly (i.e., with many errors and/or a slow rate).

The purpose of our first experiment was to investigate whether the main source of difficulty in beginning reading is at the level of connected text or at the word level. We wished to know how well one can predict a child's degree of fluency in oral reading of paragraph material from his performance (accuracy and reaction time) on selected words presented in lists.

Table 1 shows correlations between a conventional measure of fluency in oral reading, the Gray Oral Reading Test, and oral reading performance on two word lists which we devised. The Gray test consists of paragraphs of graded difficulty which yield a composite score based on time and error from which may be determined the child's reading grade level. Both word lists, which are presented as Tables 2 and 3, contain monosyllabic words. Word List 1 (Table 2) was designed primarily to study the effects of optically based ambiguity on the error pattern in reading. It consists of a number of primer words and a number of reversible words from which other words may be formed by reading from right to left. List 2 (Table 3) contains words representing equal frequencies of many of the phonemes of English and was designed specifically to make the comparison between reading and perceiving speech by ear. Data from both lists were obtained from some subjects; others received one test but not the other. Error analysis of these lists was based on phonetic transcription of the responses, and the error counts take the phoneme as the unit.[1] Our selection of this method of treating the data is explained and the procedures are described in a later section.

Table 1

Correlation of Performance of School Children on Reading Lists* and Paragraph Fluency as Measured by the Gray Oral Reading Test

Group	N	Grade	List 1	List 2
A	20	2.8	.72	--+
B	18	3.0	.77	--+
C	30	3.8	.53	.55
D	20	4.8	.77	--+

*The correlation between the two lists was .73.

+No data available.

---

[1] Our method of analysis of errors does not make any hard and fast assumptions about the size of the perceptual unit in reading. Much research on the reading process has been concerned with this problem (Huey, 1908; Woodworth, 1938; Gough, in press). Speculations have been based, for the most part, on studies of the fluent adult reader, but these studies have, nevertheless, greatly influenced theories of the acquisition of reading and views on how children should be taught (Fries, 1962; Mathews, 1966). In our view, this has had unfortunate consequences. Analysis of a well-practiced skill does not automatically reveal the stages of its acquisition, their order and special difficulties. It may be that the skilled reader does not (at all times) proceed letter by letter or even word by word, but at some stage in learning to read, the beginner probably must take account of each individual letter (Hochberg, 1970).

Table 2

Reading List 1: Containing Reversible Words, Reversible Letters, and Primer Sight Words

1. of
2. boy
3. now
4. tap
5. dog
6. lap
7. tub
8. day
9. for
10. bad
11. out
12. pat
13. ten
14. gut
15. cab
16. pit
17. saw
18. get
19. rat
20. dig
21. two
22. war
23. bed
24. felt
25. big
26. not
27. yam
28. peg
29. was
30. tab
31. won
32. pot
33. net
34. pin
35. from
36. ton
37. but
38. who
39. nip
40. on
41. bat
42. tug
43. form
44. left
45. bay
46. how
47. dip
48. no
49. pit
50. cap
51. god
52. top
53. pal
54. may
55. bet
56. raw
57. pay
58. tar
59. dab
60. tip

Table 3

Reading List 2: Presenting Equal Opportunities for Error on Each Initial Consonant,* Medial Vowel, and Final Consonant*

help
pledge
weave
lips
wreath
felt
zest
crisp
touch
palp
stash
niece
soothe
ding
that's
mesh
deep
badge
belk
gulp
stilt
zag
reach
stock
thief
coop
theme
cult
stood
these
vat
hoof
clog
move
puss
doom
talc
shoes
smooch
hook
took

teethe
stoops
bilk
hulk
jog
shook
plume
thatch
zig
teeth
moot
foot's
jeeps
leave
van
cheese
vets
loops
pooch
mash
scalp
thud
booth
wreathe
gasp
smoothe
feast
jest
chief
god
clang
dune
wasp
heath
tooth
lodge
jam
roof
gap
shove
plebe

than
dab
choose
thong
puts
hood
fun
sting
knelt
please
this
that
dub
vast
clash
soot
sheath
stop
cob
zen
push
cleave
mops
hasp
them
good
fuzz
smith
tots
such
veldt
culp
wisp
guest
bulk
silk
moose
smut
soup
fez
bing

jots
shoots
with
noose
chin
rob
plot
book
milk
vest
give
then
plug
knob
cook
love
posh
lisp
nest
sulk
zips
would
tube
chap
put
rook
loom
judge
breathe
whelp
smash
zing
could
mob
clasp
smudge
kilt
thing
fog
death
goose

thus
smelt
nudge
welt
chops
vim
vet
zip
plop
smug
foot
chest
should
clots
rasp
shops
pulp
wedge
hatch
says
watch
kelp
sheathe
bush
juice
thieve
chaff
stuff
seethe
gin
zoom
cliff
plod
rough
nook
dodge
thug
cling
news
look

---

*Consonant clusters are counted as one phoneme.

In Table 1, then, we see the correlations between the Gray Test and one or both lists for four groups of school children, all of average or above-average intelligence: Group A, 20 second grade boys (grade 2.8); Group B, 18 third grade children who comprise the lower third of their school class in reading level (grade 3.0); Group C, an entire class of 30 third grade boys and girls (grade 3.8); Group D, 20 fourth grade boys (grade 4.8).[2]

It is seen from Table 1 that for a variety of children in the early grades there is a moderate-to-high relationship between errors on the word lists and performance on the Gray paragraphs.[3] We would expect to find a degree of correlation between reading words and reading paragraphs (because the former are contained in the latter), but not correlations as high as the ones we did find if it were the case that many children could read words fluently but could not deal effectively with organized strings of words. These correlations suggest that the child may encounter his major difficulty at the level of the word--his reading of connected text tends to be only as good or as poor as his reading of individual words. Put another way, the problems of the beginning reader appear to have more to do with the synthesis of syllables than with scanning of larger chunks of connected text.

This conclusion is further supported by the results of a direct comparison of rate of scan in good- and poor-reading children by Katz and Wicklund (1971) at the University of Connecticut. Using an adaptation of the reaction-time method of Sternberg (1967), they found that both good and poor readers require 100 msec longer to scan a three-word sentence than a two-word sentence. Although, as one would expect, the poor readers were slower in reaction time than the good readers, the difference between good and poor readers remained constant as the length of the sentence was varied. (The comparison has so far been made for sentence lengths up to five words and the same result has been found: D.A. Wicklund, personal communication.) This suggests, in agreement with our findings, that good and poor readers among young children differ not in scanning rate or strategy but in their ability to deal with individual words and syllables.

As a further way of examining the relation between the rate of reading individual words and other aspects of reading performance, we obtained latency measures (reaction times) for the words in List 2 for one group of third graders (Group C, Table 1). The data show a negative correlation of .68 between latency of response and accuracy on the word list. We then compared performance on connected text (the Gray paragraphs) and on the words of List 2, and we found that latency measures and error counts showed an equal degree of (negative) correlation with paragraph reading performance. From this, it would appear that the slow rate of reading individual words may contribute as much as inaccuracy to poor performance on paragraphs. A possible explanation may be found in the rapid temporal decay in primary memory: if it takes too long to read a given word, the preceding words will have been forgotten before a phrase or sentence is completed (Gough, in press.)

---

[2] We are indebted to Charles Orlando, Pennsylvania State University, for the data in Groups A and D. These two groups comprised his subjects for a doctoral dissertation written when he was a student at the University of Connecticut (Orlando, 1971).

[3] A similarly high degree of relationship between performance on word lists and paragraphs has been an incidental finding in many studies. Jastak (1946) in his manual for the first edition of the Wide Range Achievement Test notes a correlation of .81 for his word list and the New Stanford Paragraph Reading Test. Spache (1963) cites a similar result in correlating performance on a word recognition list and paragraphs.

THE CONTRIBUTION OF VISUAL FACTORS TO THE ERROR PATTERN IN BEGINNING READING: THE PROBLEM OF REVERSALS

We have seen that a number of converging results support the belief that the primary locus of difficulty in beginning reading is the word. But, within the word, what is the nature of the difficulty? To what extent are the problems visual and to what extent linguistic?

In considering this question, we ask first whether the problem is in the perception of individual letters. There is considerable agreement that, after the first grade, even those children who have made little further progress in learning to read do not have significant difficulty in visual identification of individual letters (Vernon, 1960; Shankweiler, 1964; Doehring, 1968).

Reversals and Optical Shape Perception

The occurrence in the alphabet of reversible letters may present special problems, however. The tendency for young children to confuse letters of similar shape that differ in orientation (such as b, d, p, g, q) is well known. Gibson and her colleagues (1962; 1965) have isolated a number of component abilities in letter identification and studied their developmental course by the use of letter-like forms which incorporate basic features of the alphabet. They find that children do not readily distinguish pairs of shapes which are 180-degree transformations (i.e., reversals) of each other at age 5 or 6, but by age 7 or 8 orientation has become a distinctive property of the optical character. It is of interest, therefore, to investigate how much reversible letters contribute to the error pattern of eight-year-old children who are having reading difficulties.

Reversal of the direction of letter sequences (e.g., reading "from" for form) is another phenomenon which is usually considered to be intrinsically related to orientation reversal. Both types of reversals are often thought to be indicative of a disturbance in the visual directional scan of print in children with reading disability (see Benton, 1962, for a comprehensive review of the relevant research). One early investigator considered reversal phenomena to be so central to the problems in reading that he used the term "strephosymbolia" to designate specific reading disability (Orton, 1925). We should ask, then, whether reversals of letter orientation and sequence loom large as obstacles to learning to read. Do they co-vary in their occurrence, and what is the relative significance of the optical and linguistic components of the problem?

In an attempt to study these questions (I. Liberman, Shankweiler, Orlando, Harris, and Berti, in press) we devised the list (presented in Table 2) of 60 real-word monosyllables including most of the commonly cited reversible words and in addition a selection of words which provide ample opportunity for reversing letter orientation. Each word was printed in manuscript form on a separate 3" x 5" card. The child's task was to read each word aloud. He was encouraged to sound out the word and to guess if unsure. The responses were recorded by the examiner and also on magnetic tape. They were later analyzed for initial and final consonant errors, vowel errors, and reversals of letter sequence and orientation.

We gave List 1 twice to an entire beginning third grade class and then selected for intensive study the 18 poorest readers in the class (the lower third), because only among these did reversals occur in significant quantity.

Relationships Between Reversals and Other Types of Errors

It was found that, even among these poor readers, reversals accounted for only a small proportion of the total errors, though the list was constructed to provide maximum opportunity for reversals to occur. Separating the two types, we found that sequence reversals accounted for 15% of the total errors made and orientation errors only 10%, whereas other consonant errors accounted for 32% of the total and vowel errors 43%. Moreover, individual differences in reversal tendency were large (rates of sequence reversal ranged from 4% to 19%; rates for orientation reversal ranged from 3% to 31%). Viewed in terms of opportunities for error, orientation errors occurred less frequently than other consonant errors. Test-retest comparisons showed that whereas other reading errors were rather stable, reversals, and particularly orientation reversals, were unstable.

Reversals were not, then, a constant portion of all errors; moreover, only certain poor readers reversed appreciably, and then not consistently. Though in the poor readers we have studied, reversals are apparently not of great importance, it may be that they loom larger in importance in certain children with particularly severe and persisting reading disability. Our present data do not speak to this question. We are beginning to explore other differences between children who do and do not have reversal problems.

Orientation Reversals and Reversals of Sequence: No Common Cause?

Having considered the two types of reversals separately, we find no support for assuming that they have a common cause in children with reading problems. Among the poor third grade readers, sequence reversal and orientation reversal were found to be wholly uncorrelated with each other, whereas vowel and consonant errors correlated .73. A further indication of the lack of equivalence of the two types of reversals is that each correlated quite differently with the other error measures. It is of interest to note that sequence reversals correlated significantly with other consonant errors, with vowel errors, and with performance on the Gray paragraphs, while none of these was correlated with orientation reversals (see I. Liberman et al., in press, for a more complete account of these findings).

Orientation Errors: Visual or Phonetic?

In further pursuing the orientation errors, we examined the nature of the substitutions among the reversible letters *b*, *d*, *p* and *g*.[4] Tabulation of these showed that the possibility of generating another letter by a simple 180-degree transformation is indeed a relevant factor in producing the confusions among these letters. This is, of course, in agreement with the conclusions reached by Gibson and her colleagues (1962).

At the same time, other observations (I. Liberman et al., in press) indicated that letter reversals may be a symptom and not a cause of reading difficulty. Two observations suggest this: first, confusions among reversible letters occurred much less frequently for these same children when the letters were presented singly, even when only briefly exposed in tachistoscopic administration. If visual factors were primary, we would expect that tachistoscopic exposure would have resulted in more errors, not fewer. Secondly, the confusions among the letters during word reading were not symmetrical: as can be seen from Table 4, *b* is often confused with *p* as well as with *d*, whereas *d* tends to be confused with *b* and almost never with *p*.[5]

Table 4

Confusions Among Reversible Letters

Percentages Based on Opportunities*

Presented \ Obtained	b	d	p	g	Total Reversals	Other Errors
b	——	10.2	13.7	0.3	24.2	5.3
d	10.1	——	1.7	0.3	12.1	5.2
p	9.1	0.4	——	0.7	10.2	6.9
g	1.3	1.3	1.3	——	3.9	13.3

*Adapted from I. Liberman et al., in press.

---

[4] The letter *g* is, of course, a distinctive shape in all type styles, but it was included among the reversible letters because, historically, it has been treated as one. It indeed becomes reversible when hand printed with a straight segment below the line. Even in manuscript printing, as was used in preparing the materials for this study, the "tail" of the *g* is the only distinguishing characteristic. The letter *q* was not used because it occurs only in a stereotyped spelling pattern (*u* always following *q* in English words).

[5] The pattern of confusions among *b*, *d*, and *p* could nevertheless be explained on a visual basis. It could be argued that the greater error rate on *b* than

These findings point to the conclusion that the characteristic of optical reversibility is not a sufficient condition for the errors that are made in reading, at least among children beyond the first grade. Because the letter shapes represent segments which form part of the linguistic code, their perception differs in important ways from the perception of nonlinguistic forms--there is more to the perception of the letters in words than their shape (see Kolers, 1970, for a general discussion of this point).

Reading Reversals and Poorly Established Cerebral Dominance

S.T. Orton (1925, 1937) was one of the first to assume a causal connection between reversal tendency and cerebral ambilaterality as manifested by poorly established motor preferences. There is some clinical evidence that backward readers tend to have weak, mixed, or inconsistent hand preferences or lateral inconsistencies between the preferred hand, foot, and eye (Zangwill, 1960). Although it is doubtful that a strong case can be made for the specific association between cerebral ambilaterality and the tendency to reverse letters and letter sequences (I. Liberman et al., in press), the possibility that there is some connection between individual differences in lateralization of function and in reading disability is supported by much clinical opinion. This idea has remained controversial because, due to various difficulties, its implications could not be fully explored and tested.

It has only recently become possible to investigate the question experimentally by some means other than the determination of handedness, eyedness, and footedness. Auditory rivalry techniques provide a more satisfactory way of assessing hemispheric dominance for speech than hand preferences (Kimura, 1961; 1967).[6] We follow several investigators in the use of these dichotic

---

on _d_ or _p_ may result from the fact that _b_ offers two opportunities to make a single 180-degree transformation, whereas _d_ and _p_ offer only one. Against this interpretation we can cite further data. We had also presented to the same children a list of pronounceable nonsense syllables. Here the distribution of _b_-errors was different from that which had been obtained with real words, in that _b_ - _p_ confusions occurred only rarely. The children moreover, tended to err by converting a nonsense syllable into a word, just as in their errors on the real word lists they nearly always produced words. For this reason, a check was made of the number of real words that could be made by reversing _b_ in the two lists. This revealed no fewer opportunities to make words by substitution of _p_ than by substitution of _d_. Indeed, the reverse was the case. Such a finding lends further support to the conclusion that the nature of substitutions even among reversible letters is not an automatic consequence of the property of optical reversibility. (This conclusion was also reached by Kolers and Perkins, 1969, from a different analysis of the orientation problem.)

[6] There is reason to believe that handedness can be assessed with greater validity by substituting measures of manual dexterity for the usual questionnaire. The relation between measures of handedness and cerebral lateralization of speech, as determined by an auditory rivalry task (Shankweiler and Studdert-Kennedy, 1967), was measured by Charles Orlando (1971) in a doctoral dissertation done at the University of Connecticut. Using multiple measures of manual dexterity to assess handedness, and regarding both handedness and cerebral speech laterality as continuously distributed, Orlando found the predictive value of handedness to be high in eight- and ten-year-old children.

techniques for assessing individual differences in hemispheric specialization for speech in relation to reading ability (Kimura, personal communication; Sparrow, 1968; Zurif and Carson, 1970; Bryden, 1970). The findings of these studies as well as our own pilot work have been largely negative. It is fair to say that an association between bilateral organization of speech and poor reading has not been well supported to date.

The relationship we are seeking may well be more complex, however. Orton (1937) stressed that inconsistent lateralization for speech and motor functions is of special significance in diagnosis, and a recent finding of Bryden (1970) is of great interest in this regard. He found that boys with speech and motor functions oppositely lateralized have a significantly higher proportion of poor readers than those who show the typical uncrossed pattern. This suggests that it will be worthwhile to look closely at disparity in lateralization of speech and motor function.

If there is some relation between cerebral dominance and ability to read, we should suppose that it might appear most clearly in measures that take account not only of dominance for speech and motor function, but also of dominance for the perception of written language, and very likely with an emphasis on the relationships between them. It is known (Bryden, 1965) that alphabetical material is more often recognized correctly when presented singly to the right visual field and hence to the left cerebral hemisphere. If reliable techniques suitable for use with children can be developed for studying lateralization of component processes in reading, we suspect that much more can be learned about reading acquisition in relation to functional asymmetries of the brain.

LINGUISTIC ASPECTS OF THE ERROR PATTERN IN READING AND SPEECH

"In reading research, the deep interest in words as visual displays stands in contrast to the relative neglect of written words as linguistic units represented graphically." (Weber, 1968, p. 113)

The findings we have discussed in the preceding section suggested that the chief problems the young child encounters in reading words are beyond the stage of visual identification of letters. It therefore seemed profitable to study the error pattern from a linguistic point of view.

The Error Pattern in Misreading

We examined the error rate in reading in relation to segment position in the word (initial, medial, and final) and in relation to the type of segment (consonant or vowel).

List 2 (Table 3) was designed primarily for that purpose. It consisted of 204 real-word CVC (or CCVC and CVCC) monosyllables chosen to give equal representation to most of the consonants, consonant clusters, and vowels of English. Each of the 25 initial consonants and consonant clusters occurred eight times in the list and each final consonant or consonant cluster likewise occurred eight times. Each of eight vowels occurred approximately 25 times. This characteristic of equal opportunities for error within each consonant and vowel category enables us to assess the child's knowledge of some of the spelling patterns of English.

Table 5

Table of Phoneme Segments* Represented in the Words of List 2

Initial Consonant(s)	Vowel	Final Consonant(s)
p	a	lp
t	æ	dʒ
k	i	v
b	I	ps
d	ɛ	θ
g	ʌ	lt
m	ʊ	st
n	u	sp
w		ts
r		ʃ
l		s
f		ð
θ		ŋ
s		p
ʃ		lk
v		g
ð		tʃ
z		k
t		f
d		m
h		d
pl		z
kl		t
st		m
sm		h

*These are written in IPA.

The manner of presentation was the same as for List 1. The responses were recorded and transcribed twice by a phonetically trained person. The few discrepancies between first and second transcription were easily resolved. Although it was designed for a different purpose, List 1 also gives information about the effect of the segment position within the syllable upon error rate and the relative difficulty of different kinds of segments. We therefore analyzed results from both lists in the same way, and, as we shall see, the results are highly comparable. A list of the phoneme segments represented in the words of List 2 is shown in Table 5.

We have chosen to use phonetic transcription[7] rather than standard orthography in noting down the responses, because we believe that tabulation and analysis of oral reading errors by transcription has powerful advantages which outweigh the traditional problems associated with it. If the major sources of error in reading the words are at some linguistic level as we have argued, phonetic notation (IPA) of the responses should greatly simplify the task of detecting the sources of error and making them explicit. Transcription has the additional value of enabling us to make a direct comparison between errors in reading and in oral repetition.

Table 6 shows errors on the two word lists percentaged against opportunities as measured in four groups of school children. Group C1 includes good readers, being the upper third in reading ability of all the third graders

Table 6

Errors in Reading in Relation to Position and Type of Segment

Percentages of Opportunities for Error

Group*	Reading Ability	N	Age Range	Initial Consonant	Final Consonant	All Consonant	Vowel
C_1	Good++	11	9-10	6	12	9	10
C_2	Poor++	11	9-10	8	14	11	16
B	Poor+	18	8-10	8	14	11	27
Clinic	Poor++	10	10-12	17	24	20	31

*The groups indicated by C_1 and C_2 comprise the upper and lower thirds of Group C in Table 1. Group B is the same as so designated in Table 1. The clinic group is not represented in Table 1.

+List 1 (Table 2)

++List 2 (Table 3)

---

[7] In making the transcription, the transcriber was operating with reference to normal allophonic ranges of the phonemic categories in English.

in a particular school system; Group C2 comprises the lower third of the same third grade population mentioned above; Group B includes the lower third of the entire beginning third grade in another school system; the clinic group contains ten children, aged between 10 and 12, who had been referred to a reading clinic at the University of Connecticut. In all four groups, the responses given were usually words of English.

Table 6 shows two findings we think are important. First, there is a progression of difficulty with position of the segment in the word: final consonants are more frequently misread than initial ones; second, more errors are made on vowels than on consonants. The consistency of these findings is impressive because it transcends the particular choice of words and perhaps the level of reading ability.[8]

We will have more to say in a later section about these findings when we consider the differences between reading and speech errors. At this point, we should say that the substantially greater error rate for final consonants than for initial ones is certainly contrary to what would be expected by an analysis of the reading process in terms of sequential probabilities. If the child at the early stages of learning to read were able to utilize the constraints that are built into the language, he would take fewer errors at the end than at the beginning, not more. In fact, what we often see is that the child breaks down after he has gotten the first letter correct and can go no further. We will suggest later why this may happen.

Mishearing Differs from Misreading

In order to understand the error pattern in reading, it should be instructive to compare it with the pattern of errors generated when isolated monosyllables are presented by ear for oral repetition. We were able to make this comparison by having the same group of children repeat back a word list on one occasion and read it on another day. The ten children in the clinic group (Table 6) were asked to listen to the words in List 2 before they were asked to read them. The tape-recorded words were presented over earphones with instructions to repeat each word once. The responses were recorded on magnetic tape and transcribed in the same way as the reading responses.

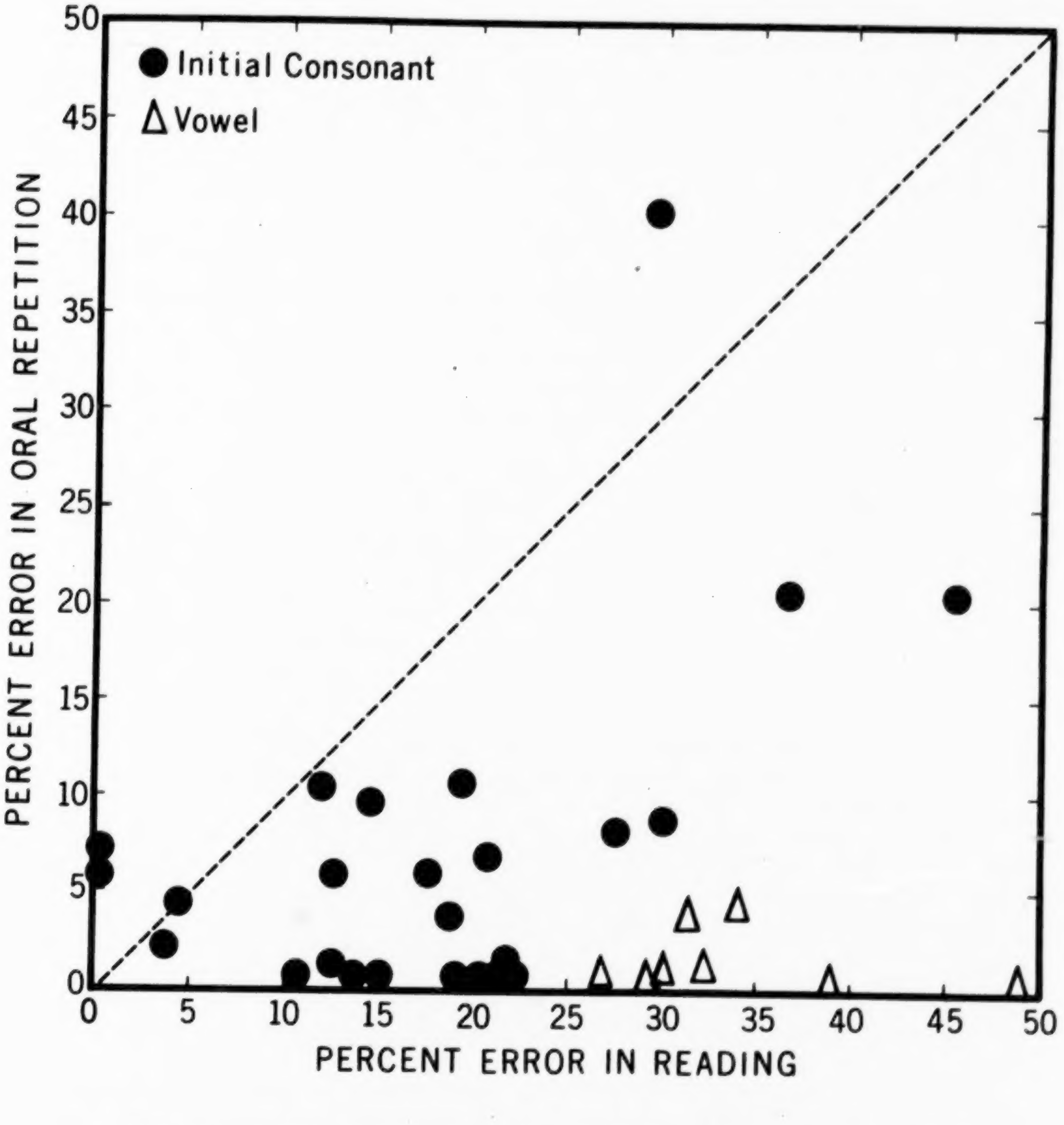
The error pattern for oral repetition shows some striking differences from that in reading. With auditory presentation, errors in oral repetition averaged 7% when tabulated by phoneme, as compared with 24% in reading, and were about equally distributed between initial and final position, rather than being markedly different. Moreover, contrary to what occurred when the list was read, fewer errors occurred on vowels than on consonants.

The relation between errors of oral repetition and reading is demonstrated in another way in the scatter plot presented as Figure 1. Percent error on initial consonants, final consonants, and vowels in reading is plotted on the abscissa against percent error on these segments in oral repetition on the ordinate. Each consonant point is based on approximately eight occurrences

---

[8] For similar findings in other research studies employing quite different reading materials and different levels of proficiency in reading, see, for example, Daniels and Diack (1956) and Weber (1970).

Scatter Diagram Showing Errors on Each Segment in Word List 2
in Relation to Opportunities



Percent error in oral repetition is plotted against percent error in reading the same words. Ten subjects.

Fig. 1

in the list over ten subjects, giving a total of 80. Each vowel point is based on approximately 25 occurrences, giving a total of 250 per point.

It is clear from the figure that the perception of speech by reading has problems which are separate and distinct from the problems of perceiving speech by ear. We cannot predict the error rate for a given phoneme in reading from its error rate in listening. If a phoneme were exactly as difficult to read as to hear, the point would fall on the diagonal line which has been dotted in. Vertical distance from the diagonal to any point below it is a measure of that phoneme's difficulty specifically in reading as distinguished from speaking and being aurally perceived. Although the reliability of the individual points in the array has not been assessed, the trends are unmistakable. The points are very widely scattered for the consonants. As for the vowels, they are seldom misheard but often misread (suggesting, incidentally, that the high error rate on vowels in reading cannot be an artifact of transcription difficulties).

Accounting for the Differences in the Error Pattern in Reading and Speech

The data presented above show that there are major differences between error patterns in reading and speech. However, they should not be taken to mean that reading and speech are not connected. What they do tell us is that reading presents special problems which reflect the difficulties of the beginning reader in making the link between segments of speech and alphabetic shapes.

Why the initial segment is more often correct in reading.

We have seen that there is much evidence to indicate that in reading the initial segment is more often correct than succeeding ones, whereas in oral repetition the error rate for initial and final consonants is essentially identical.

One of us (I. Liberman, in press) has suggested a possible explanation for this difference in distribution of errors within the syllable. She pointed out that in reading an alphabetic language like English, the child must be able to segment the words he knows into the phonemic elements which the alphabetic shapes represent. In order to do this, he needs to be consciously aware of the segmentation of the language into units of phonemic size. Seeing the word cat, being able to discriminate the individual optical shapes, being able to read the names of the three letters, and even knowing the individual sounds for the three letters cannot help him in really reading the word (as opposed to memorizing its appearance as a sight word) unless he realizes that the word in his own lexicon has three segments. Before he can map the visual message to the word in his vocabulary, he has to be consciously aware that the word cat that he knows--an apparently unitary syllable--has three separate segments. His competence in speech production and speech perception is of no direct use to him here, because this competence enables him to achieve the segmentation without ever being consciously aware of it.[9]

Though phonemic segments and their constituent features can be shown to be psychologically and physiologically real in speech perception, (A. Liberman,

---

[9] The idea of "linguistic awareness," as it has been called here, has been a recurrent theme in this conference. See especially the chapters by Ignatius Mattingly (in press) and Harris B. Savin (in press).

Cooper, Shankweiler, and Studdert-Kennedy, 1967; A. Liberman, 1968; Mattingly and Liberman, 1970), they are, as we have already noted, not necessarily available at a high level of conscious awareness. Indeed, given that the alphabetic method of writing was invented only once, and rather late in man's linguistic history, we should suspect that the phonologic elements that alphabets represent are not particularly obvious (Huey, 1908). In any event, a child whose chief problem in reading is that he cannot make explicit the phonological structure of his language might be expected to show the pattern of reading errors we found: relatively good success with the initial letters which requires no further analysis of the syllable and relatively poor performance otherwise.

Why vowel errors are more frequent in reading than in speech. Another way misreading differed from mishearing was with respect to the error rate on vowels, and we must now attempt to account for the diametrically different behavior of the vowels in reading and in oral repetition. (Of course, in the experiments we refer to here, the question is not completely separable from the question of the effect of segment position on error rate, since all vowels were medial.)

In speech, vowels, considered as acoustic signals, are more intense than consonants and they last longer. Moreover, vowel traces persist in primary memory in auditory form as "echoes." Stop consonants, on the other hand, are decoded almost immediately into an abstract phonetic form, leaving no auditory traces (Fujisaki and Kawashima, 1969; Studdert-Kennedy, 1970; Crowder, in press). At all events, one is not surprised to find that in listening to isolated words, without the benefit of further contextual cues, the consonants are most subject to error. In reading, on the other hand, the vowel is not represented by a stronger signal, vowel graphemes not being larger or more contrastful than consonant ones. Indeed, the vowels tend to suffer a disadvantage because they are usually embedded within the word. They tend, moreover, to have more complex orthographic representation than consonants.[10]

Sources of Vowel Error: Orthographic Rules or Phonetic Confusions?

The occurrence of substantially more reading errors on vowel segments than on consonant segments has been noted in a number of earlier reports (Venezky, 1968; Weber, 1970), and, as we have said, the reason usually given is that vowels are more complexly represented than consonants in English orthography. We now turn to examine the pattern of vowel errors in reading and ask what accounts for their distribution. An explanation in terms of orthography would imply that many vowel errors are traceable to misapplication of rules which involve an indirect relation between letter and sound.[11] Since the complexity of the rules varies for different vowels, it would follow that error rates among them should also vary.

---

[10] This generalization applies to English. We do not know how widely it may apply to other languages. We would greatly welcome the appearance of cross-linguistic studies of reading acquisition, which could be of much value in clarifying the relations between reading and linguistic structure. That differences among languages in orthography are related to the incidence of reading failure is often taken for granted, but we are aware of no data that directly bear on this question.

The possibility must be considered, however, that causes other than misapplication of orthographic rules may account for a larger portion of vowel misreadings. First, there could simply be a large element of randomness in the error pattern. Second, the pattern might be nonrandom, but most errors could be phonetically based rather than rule based. If reading errors on vowels have a phonetic basis, we should then expect to find the same errors occurring in reading as occur in repetition of words presented by ear. The error rate for vowels in oral repetition is much too low in our data to evaluate this possibility, but there are other ways of asking the question, as we will show.

The following analysis illustrates how vowel errors may be analyzed to discover whether, in fact, the error pattern is nonrandom and, if it is, to discover what the major substitutions are. Figure 2 shows a confusion matrix for vowels based on the responses of 11 children at the end of the third grade (Group 2 in Table 4) who are somewhat retarded in reading. Each row in the matrix refers to a vowel phoneme represented in the words (of List 2) and each column contains entries of the transcriptions of the responses given in oral reading. Thus the rows give the frequency distribution for each vowel percentaged against the number of occurrences, which is approximately 25 per vowel per subject.

It may be seen that the errors are not distributed randomly. (Chi-square computed for the matrix as a whole is 406.2 with df=42; $p < .001$). The eight vowels differ greatly in difficulty; error rates ranged from a low of 7% for /I/ to a high of 26% for /u/. Orthographic factors are the most obvious source of the differences in error rate. In our list /I/ is always represented by the letter i, whereas /u/ is represented by seven letters or digraphs: u, o, oo, ou, oe, ew, ui. The correlation (rho) between each vowel's rank difficulty and its number of orthographic representations in List 2 was .83. Hence we may conclude that the error rate on vowels in our list is related to the number of orthographic representations of each vowel.[12]

The data thus support the idea that differences in error rate among vowels reflect differences in their orthographic complexity. Moreover, as we have said, the fact that vowels, in general, map onto sound more complexly

---

[11] Some recent investigations of orthography have stressed that English spelling is more ruleful than sometimes supposed--that many seeming irregularities are actually instances of rules and that orthography operates to preserve a simpler relationship between spelling and morphophoneme at the cost of a more complex relation between spelling and sound (Chomsky and Halle, 1968; Weir and Venezky, 1968).

[12] A matrix of vowel substitutions was made up for the better readers (the upper third) of the class on which Figure 2 is based. Their distribution of errors was remarkably similar.

Matrix of Vowel Errors in Reading Word List 2, Transcribed in IPA

VOWEL OBTAINED
in Oral Reading

VOWEL PRESENTED in Print	ɑ	æ	i	ɪ	ɛ	ʌ	ʊ	u	OTHER
ɑ	87	2		1		4	1	1	4
æ	4	89		1	2	3			1
i			81	1	13				5
ɪ	1	1		93	1	3			1
ɛ	1	4	5	6	79	2	1		2
ʌ	2			3	2	80	2	4	7
ʊ	1	1				5	90	2	1
u	5	1				8	2	74	10

Each row gives the distribution of responses as percentages of opportunities for each of the eight vowels represented in the list. Eleven subjects.

Fig. 2

than consonants is one reason they tend to be misread more frequently than consonants.[13]

It may be, however, that these orthographic differences among segments are themselves partly rooted in speech. Much data from speech research indicates that vowels are often processed differently than consonants when perceived by ear. A number of experiments have shown that the tendency to categorical perception is greater in the encoded stop consonants than in the unencoded vowels (A. Liberman et al., 1967; A. Liberman, 1970). It may be argued that as a consequence of the continuous nature of their perception, vowels tend to be somewhat indefinite as phonologic entities, as illustrated by the major part they play in variation among dialects and the persistence of allophones within the same geographic locality. By the same reasoning, it could be that the continuous nature of vowel perception is one cause of complex orthography, suggesting that one reason multiple representations are tolerated may lie very close to speech.

We should also consider the possibility that the error pattern of the vowels reflects not just the complex relation between letter and sound but also confusions that arise as the reader recodes phonetically. There is now a great deal of evidence (Conrad, 1964, in press) that normal readers do, in fact, recode the letters into phonetic units for storage and use in short-term memory. If so, we should expect that vowel errors would represent displacements from the correct vowels to those that are phonetically adjacent and similar, the more so because, as we have just noted, vowel perception is more nearly continuous than categorical. That such displacements did in general occur is indicated in Figure 2 by the fact that the errors tend to lie near the diagonal. More data and, in particular, a more complete selection of items will be required to determine the contribution to vowel errors of orthographic complexity and the confusions of phonetic recoding.

SUMMARY AND CONCLUSIONS

In an attempt to understand the problems encountered by the beginning reader and children who fail to learn, we have investigated the child's misreadings and how they relate to speech. The first question we asked was whether the major barrier to achieving fluency in reading is at the level of connected text or in dealing with individual words. Having concluded from our own findings and the research of others that the word and its components are of primary importance, we then looked more closely at the error patterns in reading words.

Since reading is the perception of language by eye, it seemed important to ask whether the principal difficulties within the word are to be found at a visual stage of the process or at a subsequent linguistic stage. We considered the special case of reversals of letter sequence and orientation in which the properties of visual confusability are, on the face of it, primary. We found that although optical reversibility contributes to the error rate, it is, for the children we have studied, of secondary importance to linguistic factors. Our investigation of the reversal tendency then led us to consider whether individual differences in reading ability might reflect differences in the degree and kind of functional asymmetries of the cerebral hemispheres. Although the evidence is at this time not clearly supportive of a relation between cerebral ambilaterality and reading disability, it was suggested that new techniques offer an opportunity to explore this relationship more fully in the future.

---

[13] We did not examine consonant errors from the standpoint of individual variation in their orthographic representation, but it may be appropriate to ask whether the orthography tends to be more complex for consonants in final position than for those in initial position, since it is in the noninitial portion of words that morphophonemic alternation occurs (e.g., sign - signal). We doubt, however, that this is a major cause of the greater tendency for final consonants to be misread by beginning readers.

When we turned to the linguistic aspects of the error pattern in words, we found, as others have, that medial and final segments in the word are more often misread than initial ones and vowels more often than consonants. We then considered why the error pattern in mishearing differed from misreading in both these respects. In regard to segment position, we concluded that children in the early stages of learning to read tend to get the initial segment correct and fail on subsequent ones because they do not have the conscious awareness of phonemic segmentation needed specifically in reading but not in speaking and listening.

As for vowels in speech, we suggested, first of all, that they may tend to be heard correctly because they are carried by the strongest portion of the acoustic signal. In reading, the situation is different: alphabetic representations of the vowels possess no such special distinctiveness. Moreover, their embedded placement within the syllable and their orthographic complexity combine to create difficulties in reading. Evidence for the importance of orthographic complexity was seen in our data by the fact that the differences among vowels in error rate in reading were predictable from the number of orthographic representations of each vowel. However, we also considered the possibility that phonetic confusions may account for a significant portion of vowel errors, and we suggested how this might be tested.

We believe that the comparative study of reading and speech is of great importance for understanding how the problems of perceiving language by eye differ from the problems of perceiving it by ear and for discovering why learning to read, unlike speaking and listening, is a difficult accomplishment.

REFERENCES

Anderson, I.H. and Dearborn, W.F. (1952) The Psychology of Teaching Reading. (New York: Ronald Press).

Benton, A.L. (1962) Dyslexia in relation to form perception and directional sense. In Reading Disability, J. Money, ed. (Baltimore: Johns Hopkins Press).

Biemiller, A. (1970) The development of the use of graphic and contextual information as children learn to read. Reading Res. Quart. 6, 75-96.

Bryden, M.P. (1970) Laterality effects in dichotic listening: Relations with handedness and reading ability in children. Neuropsychologia 8, 443-450.

Bryden, M.P. (1965) Tachistoscopic recognition, handedness, and cerebral dominance. Neuropsychologia 3, 1-8.

Chomsky, N. and Halle, M. (1968) *The Sound Pattern of English*. (New York: Harper & Row).

Christenson, A. (1969) Oral reading errors of intermediate grade children at their independent, instructional, and frustration reading levels. In *Reading and Realism*, J.A. Figurel, ed., Proceedings of the International Reading Association 13, 674-677.

Conrad, R. (in press) Speech and reading. In *Language by Ear and by Eye: The Relationships between Speech and Reading*, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

Conrad, R. (1964) Acoustic confusions in immediate memory. Brit. J. Psychol. 55, 75-83.

Crowder, R. (in press) Visual and auditory memory. In *Language by Ear and by Eye: The Relationships between Speech and Reading*, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

Daniels, J.C. and Diack, H. (1956) *Progress in Reading*. (Nottingham: University of Nottingham Insitutute of Education).

Doehring, D.G. (1968) *Patterns of Impairment in Specific Reading Disability*. (Bloomington: Indiana University Press).

Fries, C.C. (1962) *Linguistics and Reading*. (New York: Holt, Rinehart and Winston).

Fujisaki, H., and Kawashima, T. (1969) On the modes and mechanisms of speech perception. *Annual Report of the Division of Electrical Engineering*, Engineering Research Institute, University of Tokyo, No. 1.

Gibson, E.J. (1965) Learning to read. Science 148, 1066-1072.

Gibson, E.J., Gibson, J.J., Pick, A.D., and Osser, R. (1962) A developmental study of the discrimination of letter-like forms. J. comp. physiol. Psychol. 55, 807-906.

Goodman, K.S. (1968) The psycholinguistic nature of the reading process. In *The Psycholinguistic Nature of the Reading Process*, K.S. Goodman, ed. (Detroit: Wayne State University Press).

Goodman, K.S. (1965) A linguistic study of cues and miscues in reading. Elementary English 42, 639-643.

Gough, P.B. (in press) One second of reading. In *Language by Ear and by Eye: The Relationships between Speech and Reading*, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

Hochberg, J. (1970) Attention in perception and reading. In *Early Experience and Visual Information Processing in Perceptual and Reading Disorders*, F.A. Young and D.B. Lindsley, eds. (Washington: National Academy of Sciences).

Huey, E.B. (1908) *The Psychology and Pedagogy of Reading*. (New York: Macmillan). (New edition, Cambridge: MIT Press, 1968.)

Jastak, J. (1946) *Wide Range Achievement Test (Examiner's Manual)*. (Wilmington, Del.: C.L. Story Co.).

Katz, L. and Wicklund, D.A. (1971) Word scanning rate for good and poor readers. J. educ. Psychol. 62, 138-140.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Kimura, D. (1961) Cerebral dominance and the perception of visual stimuli. Canad. J. of Psychol. 15, 166-171.

Kolers, P.A. (1970) Three stages of reading. In *Basic Studies on Reading*, H. Levin, ed. (New York: Harper & Row).

Kolers, P.A. and Perkins, D.N. (1969) Orientation of letters and their speed of recognition. Perception and Psychophysics 5, 275-280.

Liberman, A.M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Liberman, A.M. (1968) Discussion in *Communicating by Language: The Reading Process*, J.F. Kavanagh, ed. (Bethesda, Md.: National Institute of Child Health and Human Development) pp. 125-128.

Liberman, A.M., Cooper, F.S., Shankweiler, D., and Studdert-Kennedy, M. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, I.Y. (in press) Basic research in speech and lateralization of language: some implications for reading disability. Bull. Orton Soc. (Also in Haskins Laboratories Status Report on Speech Research 25/26, 1971, pp. 51-66.)

Liberman, I.Y., Shankweiler, D., Orlando, C., Harris, K.S., and Berti, F.B. (in press) Letter confusions and reversals of sequence in the beginning reader: Implications for Orton's theory of developmental dyslexia. Cortex. (Also in Haskins Laboratories Status Report on Speech Research 24, 1970, pp. 17-30.)

Mathews, M. (1966) *Teaching to Read Historically Considered*. (Chicago: University of Chicago Press).

Mattingly, I.G. (in press) Reading, the linguistic process and linguistic awareness. In *Language by Ear and by Eye: The Relationships between Speech and Reading*, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press). (Also in this Status Report.)

Mattingly, I.G. and Liberman, A.M. (1970) The speech code and the physiology of language. In *Information Processing in the Nervous System*, K. N. Leibovic, ed. (New York: Springer).

Orlando, C. P. (1971) Relationships between language laterality and handedness in eight and ten year old boys. Unpublished doctoral dissertation, University of Connecticut.

Orton, S.T. (1937) *Reading, Writing and Speech Problems in Children*. (New York: W.W. Norton).

Orton, S.T. (1925) "Word-blindness" in school children. Arch. Neurol. Psychiat. 14, 581-615.

Savin, H.B. (in press) What the child knows about speech when he starts to read. In *Language by Ear and by Eye: The Relationships between Speech and Reading*, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

Schale, F.C. (1966) Changes in oral reading errors at elementary and secondary levels. Unpublished doctoral dissertation, University of Chicago, 1964. (Summarized in Acad. Ther. Quart. 1, 225-229.)

Shankweiler, D. (1964) Developmental dyslexia: A critique and review of recent evidence. Cortex 1, 53-62.

Shankweiler, D. and Studdert-Kennedy, M. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. exp. Psychol. 19, 59-63.

Spache, G.D. (1963) *Diagnostic Reading Scales (Examiner's Manual)*. (Monterey, Cal.: California Test Bureau).

Sparrow, S.S. (1968) Reading disability: A neuropsychological investigation. Unpublished doctoral dissertation, University of Florida.

Sternberg, S. (1967) Two operations in character recognition: Some evidence from reaction time measures. Perception and Psychophysics 2, 45-53.

Studdert-Kennedy, M. (in press) The perception of speech. In *Current Trends in Linguistics*, Vol. XII, T.A. Sebeok, ed. (The Hague: Mouton). (Also in Haskins Laboratories Status Report on Speech Research, 23, 1970, pp. 15-48.)

Venezky, R.L. (1968) Discussion in *Communicating by Language: The Reading Process*, J.F. Kavanagh, ed. (Bethesda, Md.: National Institute of Child Health and Human Development) p. 206.

Vernon, M.D. (1960) *Backwardness in Reading*. (Cambridge: Cambridge University Press).

Weber, R. (1970) A linguistic analysis of first-grade reading errors. Reading Res. Quart. *5*, 427-451.

Weber, R. (1968) The study of oral reading errors: A survey of the literature. Reading Res. Quart. *4*, 96-119.

Weir, R.H. and Venezky, R.L. (1968) Spelling-to-sound patterns. In *The Psycholinguistic Nature of the Reading Process*, K.S. Goodman, ed. (Detroit: Wayne State University Press).

Woodworth, R.S. (1938) *Experimental Psychology*, Ch. 28 (New York: Holt).

Zangwill, O.L. (1960) *Cerebral Dominance and its Relation to Psychological Function*. (Edinburgh: Oliver & Boyd).

Zurif, E.B. and Carson, G. (1970) Dyslexia in relation to cerebral dominance and temporal analysis. Neuropsychologia *8*, 351-361.

Language Codes and Memory Codes[*]

Alvin M. Liberman,[+] Ignatius G. Mattingly,[++] and Michael T. Turvey[++]
Haskins Laboratories, New Haven

INTRODUCTION: PARAPHRASE, GRAMMATICAL CODES, AND MEMORY

When people recall linguistic information, they commonly produce utterances different in form from those originally presented. Except in special cases where the information does not exceed the immediate memory span, or where rote memory is for some reason required, recall is always a paraphrase.

There are at least two ways in which we can look at paraphrase in memory for linguistic material and linguistic episodes. We can view paraphrase as indicating the considerable degree to which detail is forgotten; at best, what is retained are several choice words with a certain syntactic structure, which, together, serve to guide and constrain subsequent attempts to reconstruct the original form of the information. On this view, rote recall is the ideal, and paraphrase is so much error. Alternatively, we can view the paraphrase not as an index of what has been forgotten but rather as an essential condition or correlate of the processes by which we normally remember. On this view, rote recall is not the ideal, and paraphrase is something other than failure to recall. It is evident that any large amount of linguistic information is not, and cannot be, stored in the form in which it was presented. Indeed, if it were, then we should probably have run out of memory space at a very early age.

We may choose, then, between two views of paraphrase: the first would say that the form of the information undergoes change because of forgetting; the second, that the processes of remembering make such change all but inevitable. In this paper we have adopted the second view, that paraphrase reflects the processes of remembering rather than those of forgetting. Putting this view another way, we should say that the ubiquitous fact of paraphrase implies that language is best transmitted in one form and stored in another.

The dual representation of linguistic information that is implied by paraphrase is important, then, if we are to store information that has been received and to transmit information that has been stored. We take it that such duality implies, in turn, a process of recoding that is somehow

---

[*] Paper presented at meeting on Coding Theory in Learning and Memory, sponsored by the Committee on Basic Research in Education, Woods Hole, Mass., August 1971.

[+] Also University of Connecticut, Storrs, and Yale University, New Haven.

[++] Also University of Connecticut, Storrs.

Acknowledgments: The authors are indebted for many useful criticisms and suggestions to Franklin S. Cooper of the Haskins Laboratories and Mark Y. Liberman of the United States Army.

constrained by a grammar. Thus, the capacity for paraphrase reflects the fundamental grammatical characteristics of language. We should say, therefore, that efficient memory for linguistic information depends, to a considerable extent, on grammar.

To illustrate this point of view, we might imagine languages that lack a significant number of the grammatical devices that all natural languages have. We should suppose that the possibilities for recoding and paraphrase would, as a consequence, be limited, and that the users of such languages would not remember linguistic information very well. Pidgins appear to be grammatically impoverished and, indeed, to permit little paraphrase, but unfortunately for our purposes, speakers of pidgins also speak some natural language, so they can convert back and forth between the natural language and the pidgin. Sign language of the deaf, on the other hand, might conceivably provide an interesting test. At the present time we know very little about the grammatical characteristics of sign language, but it may prove to have recoding (and hence paraphrase) possibilities that are, by comparison with natural languages, somewhat restricted.[1] If so, one could indeed hope to determine the effects of such restriction on the ability to remember.

In natural languages we cannot explore in that controlled way the causes and consequences of paraphrase, since all such languages must be assumed to be very similar in degree of grammatical complexity. Let us, therefore, learn what we can by looking at the several levels or representations of information that we normally find in language and at the grammatical components that convert between them.

At the one extreme is the acoustic level, where the information is in a form appropriate for transmission. As we shall see, this acoustic representation is not the whole sound as such but rather a pattern of specifiable events, the acoustic cues. By a complexly encoded connection, the acoustic cues reflect the "features" that characterize the articulatory gestures and so the phonetically distinct configurations of the vocal tract. These latter are a full level removed from the sound in the structure of language; when properly combined, they are roughly equivalent to the segments of the phonetic representation.

Only some fifteen or twenty features are needed to describe the phonetics of all human languages (Chomsky and Halle, 1968). Any particular language uses only a dozen or so features from the total ensemble, and at any particular moment in the stream of speech only six or eight features are likely to be significant. The small number of features and the complex relation between sound and feature reflect the properties of the vocal tract and the ear and also, as we will show, the mismatch between these organ systems and the requirements of the phonetic message.

At the other end of the linguistic structure is the semantic representation in which the information is ultimately stored. Because of its relative inaccessibility, we cannot speak with confidence about the shape of the information at this level, but we can be sure it is different from the acoustic. We should suppose, as many students do, that the semantic information is also to be described in terms of features. But if the indefinitely many aspects of experience are to be represented, then the available inventory of semantic features must be very large, much larger surely than the dozen or so phonetic features that will be used as the ultimate vehicles. Though particular semantic sets may comprise many features, it is conceivable that the structure of a set might be quite simple. At all events, the characteristics of the semantic representation can be assumed to reflect properties of long-term memory, just as the very different characteristics of the acoustic and phonetic representations reflect the properties of components most directly concerned with transmission.

---

[1] The possibilities for paraphrase in sign language are, in fact, being investigated by Edward Klima and Ursula Bellugi.

The gap between the acoustic and semantic levels is bridged by grammar. But the conversion from the one level to the other is not accomplished in a single step, nor is it done in a simple way. Let us illustrate the point with a view of language like the one developed by the generative grammarians (see Chomsky, 1965). On that view there are three levels--deep structure, surface structure, and phonetic representation--in addition to the two--acoustic and semantic--we have already talked about. As in the distinction between acoustic and semantic levels, the information at every level has a different structure. At the level of deep structure, for example, a string such as The man sings. The man married the girl. The girl is pretty. becomes at the surface The man who sings married the pretty girl. The restructuring from one level to the next is governed by the appropriate component of the grammar. Thus, the five levels or streams of information we have identified would be connected by four sets of grammatical rules: from deep structure to the semantic level by the semantic rules; in the other direction, to surface structure, by syntactic rules; then to phonetic representation by phonologic rules; and finally to the acoustic signal by the rules of speech.[2] It should be emphasized that none of these conversions is straightforward or trivial, requiring only the substitution of one segment or representation for another. Nor is it simply a matter of putting segments together to form larger units, as in the organization of words into phrases and sentences or of phonetic segments into syllables and breath groups. Rather, each grammatical conversion is a true restructuring of the information in which the number of segments, and often their order, is changed, sometimes drastically. In the context of the conference for which this paper was prepared, it is appropriate to describe the conversions from one linguistic level to another as recodings and to speak of the grammatical rules which govern them as codes.

Paraphrase of the kind we implied in our opening remarks would presumably occur most freely in the syntactic and semantic codes. But the speech code, at the other end of the linguistic structure, also provides for a kind of paraphrase. At all events it is, as we hope to show, an essential component of the process that makes possible the more obvious forms of paraphrase, as well as the efficient memory which they always accompany.

---

[2] In generative grammar, as in all others, the conversion between phonetic representation and acoustic signal is not presumed to be grammatical. As we have argued elsewhere, however, and as will to some extent become apparent in this paper, this conversion is a complex recoding, similar in formal characteristics to the recodings of syntax and phonology (Mattingly and Liberman, 1969; Liberman, 1970).

Grammar is, then, a set of complex codes that relates transmitted sound and stored meaning. It also suggests what it is that the recoding processes must somehow accomplish. Looking at these processes from the speaker's viewpoint, we see, for example, that the semantic features must be replaced by phonological features in preparation for transmission. In this conversion an utterance which is, at the semantic level, a single unit comprising many features of meaning becomes, phonologically, a number of units composed of a very few features, the phonologic units and features being in themselves meaningless. Again, the semantic representation of an utterance in coherent discourse will typically contain multiple references to the same topic. This amounts to a kind of redundancy which serves, perhaps, to protect the semantic representation from noise in long-term memory. In the acoustic representation, however, to preserve such repetitions would unduly prolong discourse. To take again the example we used earlier, we do not say The man sings. The man married the girl. The girl is pretty. but rather The man who sings married the pretty girl. The syntactic rules describe the ways in which such redundant references are deleted. At the acoustic and phonetic levels, redundancy of a very different kind may be desirable. Given the long strings of empty elements that exist there, the rules of the phonologic component predict certain lawful phonetic patterns in particular contexts and, by this kind of redundancy, help to keep the phonetic events in their proper order.

But our present knowledge of the grammar does not provide much more than a general framework within which to think about the problem of recoding in memory. It does not, for example, deal directly with the central problem of paraphrase. If a speaker-hearer has gone from sound to meaning by some set of grammatical rules, what is to prevent his going in the opposite direction by the inverse operations, thus producing a rote rendition of the originally presented information? In this connection we should say on behalf of the grammar that it is not an algorithm for automatically recoding in one direction or the other, but rather a description of the relationships that must hold between the semantic representation, at the one end, and the corresponding acoustic representation at the other. To account for paraphrase, we must suppose that the speaker synthesizes the acoustic representation, given the corresponding semantic representation, while the listener must synthesize an approximately equivalent semantic representation, given the corresponding acoustic representation. Because the grammar only constrains these acts of synthesis in very general ways, there is considerable freedom in the actual process of recoding; we assume that such freedom is essential if linguistic information is to be well remembered.

For students of memory, grammatical codes are unsatisfactory in yet another, if closely related, respect: though they may account for an otherwise arbitrary-appearing relation between streams of information at different levels of the linguistic structure, they do not describe the actual processes by which the human being recodes from the one level to the other, nor does the grammarian intend that they should. Indeed, it is an open question whether even the levels that the grammar assumes--for example, deep structure--have counterparts of some kind in the recoding process.

We might do well, then, to concentrate our attention on just one aspect of grammar, the speech code that relates the acoustic and phonetic representations, because we may then avoid some of the difficulties we encounter in the "higher" or "deeper" reaches of the language. The acoustic and phonetic levels have been accessible to psychological (and physiological) experiment, as a result of which we are able to talk about "real" processes and "real" levels, yet the conversion we find there resembles grammatical codes more generally and can be shown, in a functional as well as a formal sense, to be an integral part of language. We will, therefore, examine in some detail the characteristics of the speech code, having in mind that it reflects some of the important characteristics of the broader class of language codes and that it may, therefore, serve well as a basis for comparison with the memory codes we are supposed to be concerned with. It is the more appropriate that we should deal with the speech code because it comprises the conversion from an acoustic signal appropriate for transmission to a phonetic representation appropriate for storage in short-term memory, a process that is itself of some interest to members of this conference.

CHARACTERISTICS OF THE SPEECH CODE

Clarity of the Signal

It is an interesting and important fact about the speech code that the physical signal is a poor one. We can see that this is so by looking at a spectrographic representation of the speech signal like the one in Figure 1. This is a picture of the phrase "to catch pink salmon." As always in a spectrogram, frequency is on the vertical axis, time on the horizontal; relative intensity is represented by the density, or blackness, of the marks. The relatively darker bands are resonances of the vocal tract, the so-called formants. We know that the lowest two or three of these formants contain almost all of the linguistic information; yet, as we can see, the acoustic energy is not narrowly concentrated there but tends rather to be smeared across the spectrum; moreover, there is at least one higher formant at about 3600 cps that never varies and thus carries no linguistic information at all. This is to say that the linguistically important cues constitute a relatively small part of the total physical energy. To appreciate to what extent this is so, we might contrast speech with the printed alphabet, where the important parts of the signal stand out clearly from the background. We might also contrast a spectrogram of the "real" speech of Figure 1 with a "synthetic" spectrogram like the one in Figure 2, which produces intelligible speech though the formants are unnaturally narrow and sharply defined.

In fact, the speech signal is worse than we have so far said or than we can immediately see just by looking at a spectrogram, for, paradoxically, the formants are most indeterminate at precisely those points where the information they carry is most important. It is, we know, the rapid changes in the frequency position of the formants (the formant transitions) that contain the essential cues for most of the consonants. In the case of the stop consonants, these changes occur in 50 msec or less, and they sometimes extend over ranges as great as 600 cps. Such signals scatter energy and are therefore difficult to specify or to track. Moreover, the difficulty is greatest at the point where they begin, though that is the most important part of the transition for the listener who wants to know the phonetic identity of sound.

Spectrogram of "to catch pink salmon," Natural Speech

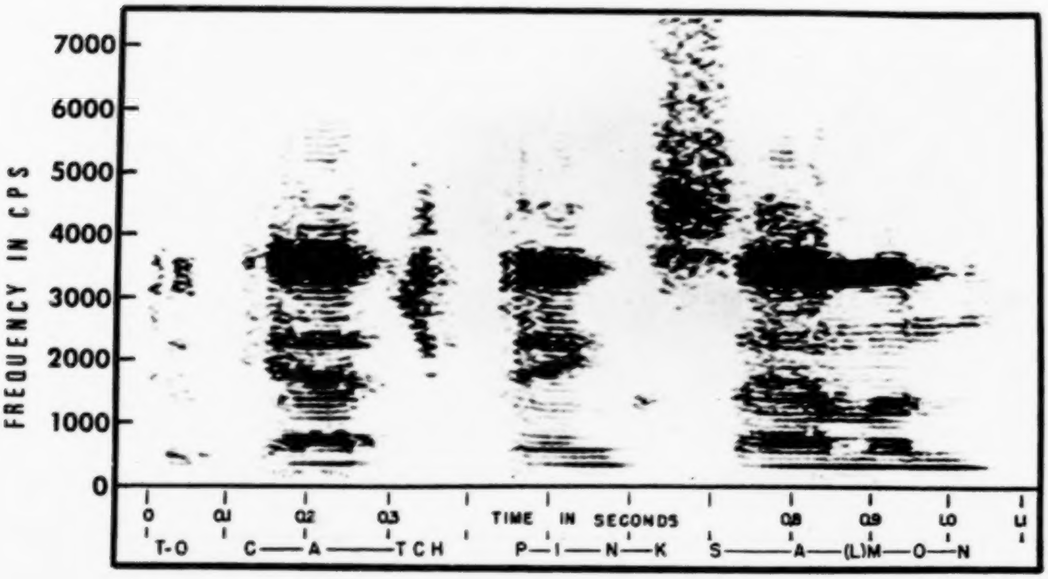


Fig. 1

Schematic Spectrogram for Synthesis of "to catch pink salmon"

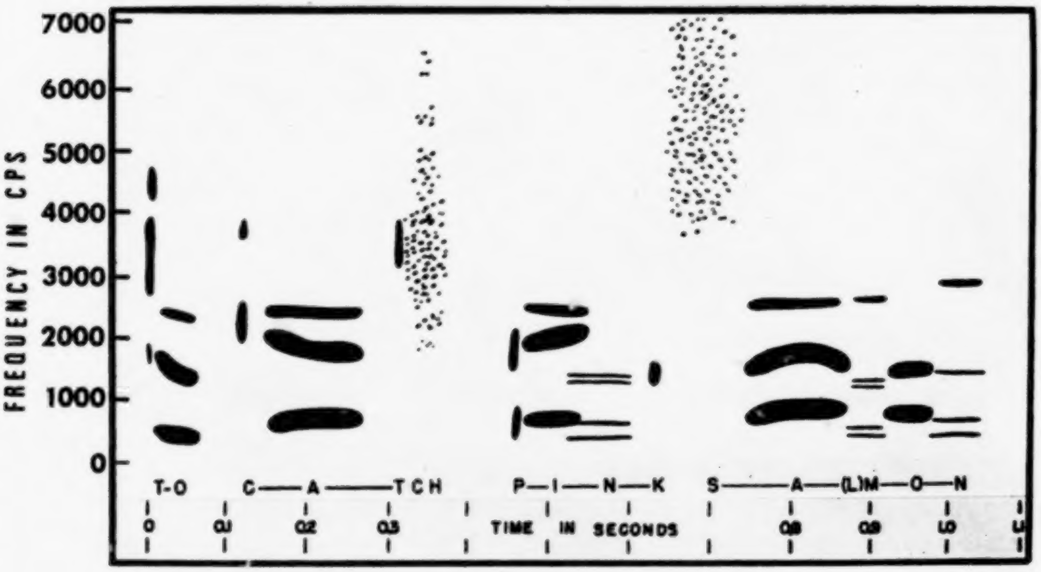


Fig. 2

The physical indeterminacy of the signal is an interesting aspect of the speech code because it implies a need for processors specialized for the purpose of extracting the essential acoustic parameters. The output of these processors might be a cleaned-up description of the signal, not unlike the simplified synthetic spectrogram of Figure 2. But such an output, it is important to understand, would be auditory, not phonetic. The signal would only have been clarified; it would not have been decoded.

Complexity of the Code

Like the other parts of the grammatical code, the conversion from speech sound to phonetic message is complex. Invoking a distinction we have previously found useful in this connection, we should say that the conversion is truly a code and not a cipher (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Studdert-Kennedy, in press). If the sounds of speech were a simple cipher, there would be a unit sound for each phonetic segment. Something approximating such a cipher does indeed exist in one of the written forms of language--viz., alphabets--where each phonological[3] segment is represented by a discrete optical shape. But speech is not an alphabet or cipher in that sense. In the interconversion between acoustic signal and phonetic message the information is radically restructured so that successive segments of the message are carried simultaneously--that is, in parallel--on exactly the same parts of the acoustic signal. As a result, the segmentation of the signal does not correspond to the segmentation of the message; and the part of the acoustic signal that carries information about a particular phonetic segment varies drastically in shape according to context.

In Figure 3 we see schematic spectrograms that produce the syllables [di] and [du] and illustrate several aspects of the speech code. To synthesize the vowels [i] and [u], at least in slow articulation, we need only the steady-state formants--that is, the parts of the pattern to the right of the formant transitions. These acoustic segments correspond in simple fashion to the perceived phonetic segments: they provide sufficient cues for the vowels; they carry information about no other segments; and though the fact is not illustrated here, they are, in slow articulation, the same in all message contexts. For the slowly articulated vowels, then, the relation between sound and message is a simple cipher. The stop consonants, on the other hand, are complexly encoded, even in slow articulation. To see in what sense this is so, we should examine the formant transitions, the rapid changes in formant frequency at the beginning (left) of the pattern. Transitions of the first (lower) formant are cues for manner and voicing; in this case they tell the listener that the consonants are members of the class of voiced stops [bdg]. For our present purposes, the transitions of the second (higher) formant--the parts of the pattern enclosed in the broken circles--are of greater interest. Such transitions are, in general, cues for the perceived "place" distinctions

---

[3] Alphabets commonly make contact with the language at a level somewhat more abstract than the phonetic. Thus, in English the letters often represent what some linguists would call morphophonemes, as for example in the use of "s" for what is phonetically the [s] of cats and the [z] of dogs. In the terminology of generative grammar, the level so represented corresponds roughly to the phonological.

Schematic Spectrogram for the Syllables [di] and [du]

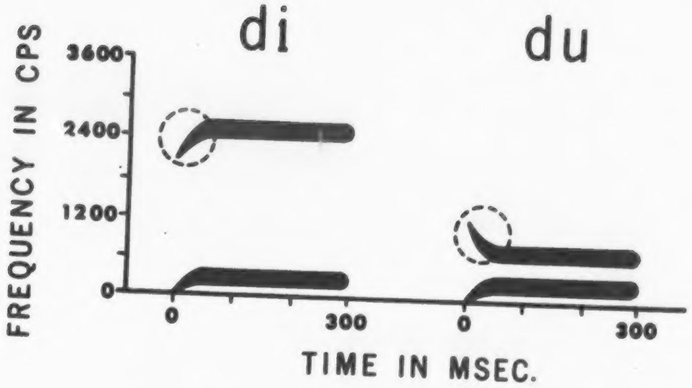


Fig. 3

among the consonants. In the patterns of Figure 3 they tell the listener that the stop is [d] in both cases. Plainly, the transition cues for [d] are very different in the two vowel contexts: the one with [i] is a rising transition relatively high in the spectrum, the one with [u] a falling transition low in the spectrum. It is less obvious, perhaps, but equally true that there is no isolable acoustic segment corresponding to the message segment [d]: at every instant, the second-formant transition carries information about both the consonant and the vowel. This kind of parallel transmission reflects the fact that the consonant is truly encoded into the vowel; this is, we would emphasize, the central characteristic of the speech code.

The next figure (Figure 4) shows more clearly than the last the more complex kind of parallel transmission that frequently occurs in speech. If converted to sound, the schematic spectrogram shown there is sufficient to produce an approximation to the syllable [bæg]. The point of the figure is to show where information about the phonetic segments is to be found in the acoustic signal. Limiting our attention again to the second formant, we see that information about the vowel extends from the beginning of the utterance to the end. This is so because a change in the vowel--from [bæg] to [big], for example--will require a change in the entire formant, not merely somewhere in its middle section. Information about the first consonant, [b], extends through the first two-thirds of the whole temporal extent of the formant. This can be established by showing that a change in the first segment of the message--from [bæg] to [gæg], for example--will require a change in the signal from the beginning of the sound to the point, approximately two-thirds of the way along the formant, that we see marked in the figure. A similar statement and similar test apply also to the last consonant, [g]. In general, every part of the second formant carries information about at least two segments of the message; and there is a part of that formant, in the middle, into which all three message segments have been simultaneously encoded. We see, perhaps more easily than in Figure 1, that the lack of correspondence in segmentation is not trivial. It is not the case that there are simple extensions connecting an otherwise segmented signal, as in the case of cursive writing, or that there are regions of acoustic overlap separating acoustic sections that at some point correspond to the segments of the message. There is no correspondence in segmentation because several segments of the message have been, in a very strict sense, encoded into the same segment of the signal.

Transparency of the Code

We have just seen that not all phonetic segments are necessarily encoded in the speech signal to the same degree. In even the slowest articulations, all of the consonants, except the fricatives,[4] are encoded. But the vowels (and the fricatives) can be, and sometimes are, represented in the acoustic signal quite straightforwardly, one acoustic segment for each phonetic segment. It is as if there were in the speech stream occasionally transparent stretches. We might expect that these stretches, in which the phonetic elements are not restructured in the sound, could be treated as if they were a

---

[4] For a fuller discussion of this point, see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967.

Schematic Spectrogram Showing Effects of Coarticulation in the Syllable [bæg]

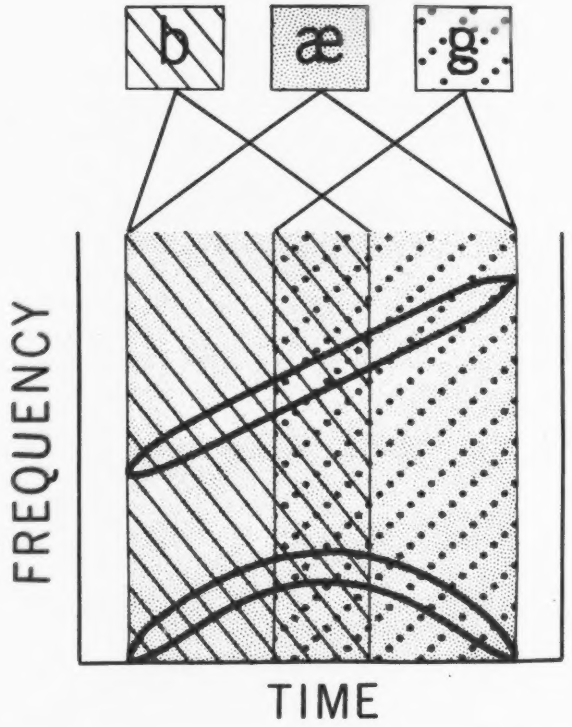


Fig. 4

cipher. There is, thus, a kind of intermittency in the difficulty of decoding the acoustic signal. We may wonder whether that characteristic of the speech code serves a significant purpose--such as providing the decoding machinery with frequent opportunities to get back on the track when and if things go wrong--but it is, in any case, an important characteristic to note, as we will see later in the paper, because of the correspondence between what we might call degree of encoding and evidence for special processing.

Lawfulness of the Code

Given an encoded relation between two streams or levels of information such as we described in the preceding section, we should ask whether the conversion from the one to the other is made lawfully--that is, by the application of rules--or, alternatively, in some purely arbitrary way. To say that the conversion is by rule is to say that it can be rationalized, that there is, in linguistic terms, a grammar. If the connection is arbitrary, then there is, in effect, a code book; to decode a signal, one looks it up in the book.

The speech code is, as we will see, not arbitrary, yet it might appear so to an intelligent but inarticulate cryptanalyst from Mars. Suppose that such a creature, knowing nothing about speech, were given many samples of utterances (in acoustic or visible form), each paired with its decoded or plain-text phonetic equivalents. Let us suppose further, as seems to us quite reasonable, that he would finally conclude that the code could not be rationalized, that it could only be dealt with by reference to a code book. Such a conclusion would, of course, be uninteresting. From the point of view of one who knows that human beings readily decode spoken utterances, the code-book solution would also seem implausible, since the number of entries in the book would have to be so very large. Having in mind the example of [bæg] that we developed earlier, we see that the number of entries would, at the least, be as great as the number of syllables. But, in fact, the number would be very much larger than that, because coding influences sometimes extend across syllable boundaries (Ohman, 1966) and because the acoustic shape of the signal changes drastically with such factors as rate of speaking and phonetic stress (Lindblom, 1963; Lisker and Abramson, 1967).

At all events, our Martian would surely have concluded, to the contrary, that the speech code was lawful if anyone had described for him, even in the most general terms, the processes by which the sounds are produced. Taking the syllable [bæg], which we illustrated earlier, as our example, one might have offered a description about as follows. The phonetic segments of the syllable are taken apart into their constituent features, such as place of production, manner of production, condition of voicing, etc. These features are represented, we must suppose, as neural signals that will become, ultimately, the commands to the muscles of articulation. Before they become the final commands, however, the neural signals are organized so as to produce the greatest possible overlap in activity of the independent muscles to which the separate features are assigned. There may also occur at this stage some reorganization of the commands so as to insure cooperative activity of the several muscle groups, especially when they all act on the same organ, as is the case with the muscle groups that control the gestures of the tongue. But so far the features, or rather their neural equivalents, have only been organized; they can still be found as largely independent entities, which is to say that they have not yet been thoroughly encoded. In the next stage the neural commands (in the final common paths) cause muscular contraction, but this conversion is, from our standpoint, straightforward and need not detain us. It is in the final conversions, from muscle contraction to vocal-tract shape to sound, that the output is radically restructured and that true encoding occurs. For it is there that the independent but overlapping activity of independent muscle groups becomes merged as they are reflected in the acoustic signal. In the case of [bæg], the movement of the lips that represents a feature of the initial consonant is overlapped with the shaping of the tongue appropriate for the next vowel segment. In the conversion to sound, the number of dimensions is reduced, with the result that the simultaneous activity of lips and tongue affect exactly the same parameter of the acoustic signal, for example, the second formant. We, and our Martian, see then how it is that the consonant and the vowel are encoded.

The foregoing account is intended merely to show that a very crude model can, in general, account for the complexly encoded relation between the speech signal and the phonetic message. That model rationalizes the relation between these two levels of the language, much as the linguists' syntactic model rationalizes the relation between deep and surface structure. For that reason, and because of certain formal similarities we have described elsewhere (Mattingly and Liberman, 1969), we should say of our speech model that it is, like syntax, a grammar. It differs from syntax in that the grammar of speech is a model of a flesh-and-blood process, not, as in the case of syntax, a set of rules with no describable physiological correlates. Because the grammar of speech corresponds to an actual process, we are led to believe that it is important, not just to the scientist who would understand the code but also to the ordinary listener who needs that same kind of understanding, albeit tacitly, if he is to perform appropriately the complex task of perceiving speech. We assume that the listener decodes the speech signal by reference to the grammar, that is, by reference to a general model of the articulatory process. This assumption has been called the motor theory of speech perception.

Efficiency of the Code

The complexity of the speech code is not a fluke of nature that man has somehow got to cope with but is rather an essential condition for the efficiency of speech, both in production and in perception, serving as a necessary link between an acoustic representation appropriate for transmission and a phonetic representation appropriate for storage in short-term memory. Consider production first. As we have already had occasion to say, the constituent features of the phonetic segments are assigned to more or less independent sets of articulators, whose activity is then overlapped to a very great extent. In the most extreme case, all the muscle movements required to communicate the entire syllable would occur simultaneously; in the more usual case, the activity corresponding to the several features is broadly smeared through the syllable. In either case the result is that phonetic segments are realized in articulation at rates higher than the rate at which any single muscle can change its state. The coarticulation that characterizes so much of speech production and causes the complications of the speech code seems well designed to permit relatively slow-moving muscles to transmit phonetic segments at high rates (Cooper, 1966).

The efficiency of the code on the side of perception is equally clear. Consider, first, that the temporal resolving power of the ear must set an upper limit on the rate at which we can perceive successive acoustic events. Beyond that limit the successive sounds merge into a buzz and become unidentifiable. If speech were a cipher on the phonetic message--that is, if each segment of the message were represented by a unit sound--then the limit would be determined directly by the rate at which the phonetic segments were transmitted. But given that the message segments are, in fact, encoded into acoustic segments of roughly syllabic size, the limit is set not by the number of phonetic segments per unit time but by the number of syllables. This represents a considerable gain in the rate at which message segments can be perceived.

The efficient encoding described above results from a kind of parallel transmission in which information about successive segments is transmitted simultaneously on the same part of the signal. We should note that there is another, very different kind of parallel transmission in speech: cues for the features of the same segment are carried simultaneously on different parts of the signal. Recalling the patterns of Figure 4, we note that the cues for place of production are in the second-formant transition, while the first-formant transition carries the cues for manner and voicing. This is an apparently less complicated arrangement than the parallel transmission produced by the encoding of the consonant into the vowel, because it takes advantage of the ear's ability to resolve two very different frequency levels. We should point out, however, that the listener is not at all aware of the two frequency levels, as he is in listening to a chord that is made up of two pitches, but rather hears the stop, with all its features, in a unitary way.

The speech code is apparently designed to increase efficiency in yet another aspect of speech perception: it makes possible a considerable gain in our ability to identify the order in which the message segments occur. Recent research by Warren et al. (1969) has shown that the sequential order of nonspeech signals can be correctly identified only when these segments have durations several times greater than the average that must be assigned to the message segments in speech. If speech were a cipher--that is, if there were an invariant sound for each unit of the message--then it would have to be transmitted at relatively low rates if we were to know that the word "task," for example, was not "taks" or "sakt" or "kats." But in the speech code, the order of the segments is not necessarily signalled, as we might suppose, by the temporal order in which the acoustic cues occur. Recalling what we said earlier about the context-conditioned variation in the cues, we should note now that each acoustic cue is clearly marked by these variations for the position of the signalled segment in the message. In the case of the transition cues for [d] that we described earlier, for example, we should find that in initial and final positions--for example, in [dæg] and [gæd]--the cues were mirror images. In listening to speech we somehow hear through the context-conditioned variation in order to arrive at the canonical form of the segment, in this case [d]. But we might guess that we also use the context-determined shape of the cue to decide where in the sequence the signalled segment occurred. In any case, the order of the segments we hear may be to a large extent inferred--quite exactly synthesized, created, or constructed--from cues in a way that has little or nothing to do with the order of their occurrence in time. Given what appears to be a relatively poor ability to identify the order of acoustic events from temporal cues, this aspect of the speech code would significantly increase the rate at which we can accurately perceive the message.

The speech code is efficient, too, in that it converts between a high-information-cost acoustic signal appropriate for transmission and a low-information-cost phonetic string appropriate for storage in some short-term memory. Indeed, the difference in information rate between the two levels of the speech code is staggering. To transmit the signal in acoustic form and in high fidelity costs about 70,000 bits per second; for reasonable intelligibility we need about 40,000 bits per second. Assuming a frequency-volley theory of hearing through most of the speech range, we should suppose that a great deal of nervous tissue would have to be devoted to the storage of even relatively short stretches. But recoding into a phonetic representation, we reduce the cost to less than 40 bits per second, thus effecting a saving of about 1,000 times by comparison with the acoustic form and of roughly half that by comparison with what we might assume a reduced auditory (but not phonetic) representation to be. We must emphasize, however, that this large saving is realized only if each phonetic feature is represented by a unitary pattern of nervous activity, one such pattern for each feature, with no additional or extraneous "auditory" information clinging to the edges. As we will see in the next section, the highly encoded aspects of speech do tend to become highly digitized in that sense.

Naturalness of the Code

It is testimony to the naturalness of the speech code that all members of our species acquire it readily and use it with ease. While it is surely true that a child reared in total isolation would not produce phonetically intelligible speech, it is equally true that in normal circumstances he comes to do that without formal tuition. Indeed, given a normal child in a normal environment, it would be difficult to contrive methods that would effectively prevent him from acquiring speech.

It is also relevant that, as we pointed out earlier, there is a universal phonetics. A relatively few phonetic features suffice, given the various combinations into which they are entered, to account for most of the phonetic segments, and in particular those that carry the heaviest information load, in the languages of the world. For example, stops and vowels, the segments with which we have been exclusively concerned in this paper, are universal, as is the co-articulated consonant-vowel syllable that we have used to illustrate the speech code. Such phonetic universals are the more interesting because they often require precise control of articulation; hence they are not to be dismissed with the airy observation that since all men have similar vocal tracts, they can be expected to make similar noises.

Because the speech code is complex but easy, we should suppose that man has access to special devices for encoding and decoding it. There is now a great deal of evidence that such specialized processors do exist in man, apparently by virtue of his membership in the race. As a consequence, speech requires no conscious or special effort; the speech code is well matched to man and is, in precisely that sense, natural.

The existence of special speech processors is strongly suggested by the fact that the encoded sounds of speech are perceived in a special mode. It is obvious--indeed so obvious that everyone takes it for granted--that we do not and cannot hear the encoded parts of the speech signal in auditory terms. The first segment of the syllables [ba], [da], [ga] have no identifiable auditory characteristics; they are unique linguistic events. It is as if they were the abstract output of a device specialized to extract them, and only them, from the acoustic signal. This abstract nonauditory perception is characteristic of encoded speech, not of a class of acoustic events such as the second-formant transitions that are sufficient to distinguish [ba], [da], [ga], for when these transition cues are extracted from synthetic speech patterns and presented alone, they sound just like the "chirps" or glissandi that auditory psychophysics would lead us to expect. Nor is this abstract perception characteristic of the relatively unencoded parts of the speech signal: the steady-state noises of the fricatives, [s] and [ʃ], for example, can be heard as noises; moreover, one can easily judge that the noise of [s] is higher in pitch than the noise of [ʃ].

A corollary characteristic of this kind of abstract perception, measured quite carefully by a variety of techniques, is one that has been called "categorical perception" (see Studdert-Kennedy, Liberman, Harris, and Cooper, 1970, for a review; Haggard, 1970, 1971b; Pisoni, 1971; Vinegrad, 1970). In listening to the encoded segments of speech we tend to hear them only as categories, not as a perceived continuum that can be more or less arbitrarily divided into regions. This occurs even when, with synthetic speech, we produce stimuli that lie at intermediate points along the acoustic continuum that contains the relevant cues. In its extreme form, which is rather closely approximated in the case of the stops, categorical perception creates a situation, very different from the usual psychophysical case, in which the listener can discriminate stimuli as different no better than he can identify them absolutely.

That the categorical perception of the stops is not simply a characteristic of the way we process a certain class of acoustic stimuli--in this case the rapid frequency modulation that constitutes the (second-formant transition) acoustic cue--has been shown in a recent study (Mattingly, Liberman, Syrdal, and Halwes, 1971). It was found there that, when listened to in isolation, the second-formant transitions--the chirps we referred to earlier--are not perceived categorically.

Nor can it be said that categorical perception is simply a consequence of our tendency to attach phonetic labels to the elements of speech and then to forget what the elements sounded like. If that were the case, we should expect to find categorical perception of the unencoded steady-state vowels, but in fact, we do not--certainly not to the same extent (Fry, Abramson, Eimas, and Liberman, 1962; Eimas, 1963; Stevens, Liberman, Ohman, and Studdert-Kennedy, 1969; Pisoni, 1971; Fujisaki and Kawashima, 1969). Moreover, categorical perception of the encoded segments has recently been found to be reflected within 100 msec in cortical evoked potentials (Dorman, 1971).

In the case of the encoded stops, then, it appears that the listener has no auditory image of the signal available to him, but only the output of a specialized processor that has stripped the signal of all normal sensory information and represented each phonetic segment (or feature) categorically by a unitary neural event. Such unitary neural representations would presumably be easy to store and also to combine, permute, and otherwise shuffle around in the further processing that converts between sound and meaning.

But perception of vowels is, as we noted, not so nearly categorical. The listener discriminates many more stimuli than he can absolutely identify, just as he does with nonspeech; accordingly, we should suppose that, as with nonspeech, he hears the signal in auditory terms. Such an auditory image would be important in the perception of the pitch and duration cues that figure in the prosodic aspects of speech; moreover, it would be essential that the auditory image be held for some seconds, since the listener must often wait to the end of a phrase or sentence in order to know what linguistic value to assign to the particular pitch and duration cues he heard earlier.

Finally, we should note about categorical perception that, according to a recent study (Eimas, Siqueland, Jusczyk, and Vigorito, 1971), it is present in infants at the age of four weeks. These infants discriminated synthetic [ba] and [pa]; moreover, and more significantly, they discriminated better, other things being equal, between pairs of stimuli which straddled the adult phonetic boundary than between pairs which lay entirely within the phonetic category. In other words, the infants perceived the voicing feature categorically. From this we should conclude that the voicing feature is real, not only physiologically but in a very natural sense.

Other, perhaps more direct, evidence for the existence of specialized speech processors comes from a number of recent experiments that overload perceptual mechanisms by putting competing signals simultaneously into the two ears (Broadbent and Gregory, 1964; Bryden, 1963; Kimura, 1961, 1964, 1967; Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970). The general finding with speech signals, including nonsense syllables that differ, say, only in the initial consonant, is that stimuli presented to the right ear are better heard than those presented to the left; with complex nonspeech sounds the opposite result--a left-ear advantage--is found. Since there is reason to believe, especially in the case of competing and dichotically presented stimuli, that the contralateral cerebral representation is the stronger, these results have been taken to mean that speech, including its purely phonetic aspects, needs to be processed in the left hemisphere, nonspeech in the right. The fact that phonetic perception goes on in a particular part of the brain is surely consistent with the view that it is carried out by a special processor.

The case for a special processor to decode speech is considerably strengthened by the finding that the right-ear advantage depends on the encodedness of the signal. For example, stop consonants typically show a larger and more consistent right-ear advantage than unencoded vowels (Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970). Other recent studies have confirmed that finding and have explored even more analytically the conditions of the right-ear (left-hemisphere) advantage for speech (Darwin, 1969, 1971; Haggard, 1971a; Haggard, Ambler, and Callow, 1969; Haggard and Parkinson, 1971; Kirstein and Shankweiler, 1969; Spellacy and Blumstein, 1970). The results, which are too numerous and complicated to present here even in summary form, tend to support the conclusion that processing is forced into the left hemisphere (for most subjects) when phonetic decoding, as contrasted with phonetic deciphering or with processing of nonspeech, must be carried out.

Having referred in the discussion of categorical perception to the evidence that the phonetic segments (or, rather, their features) may be assumed to be represented by unitary neural events, we should here point to an incidental result of the dichotic experiments that is very relevant to that assumption. In three experiments (Halwes, 1969; Studdert-Kennedy and Shankweiler, 1970; Yoder, pers. comm.) it has been found that listeners tend significantly often to extract one feature (e.g., place of production) from the input to one ear and another feature (e.g., voicing) from the other and combine them to hear a segment that was not presented to either ear. Thus, given [ba] to the left ear, say, and [ka] to the right, listeners will, when they err, far more often report [pa] (place feature from the left ear, voicing from the right) or [ga] (place feature from the right ear, voicing from the left) than [da] or [ta]. We take this as conclusive evidence that the features are singular and unitary in the sense that they are independent of the context in which they occur and also that, far from being abstract inventions of the linguist, they have, in fact, a hard reality in physiological and psychological processes.

The technique of overloading the perceptual machinery by dichotic presentation has led to the discovery of yet another effect which seems, so far, to testify to the existence of a special speech processor (Studdert-Kennedy, Shankweiler, and Schulman, 1970). The finding, a kind of backward masking that has been called the "lag" effect, is that when syllables contrasting in the initial stop consonant are presented dichotically and offset in time, the second (or lagging) syllable is more accurately perceived. When such syllables are presented monotically, the first (or leading) stimulus has the advantage. In the dichotic case, the effect is surely central; in the monotic case there is presumably a large peripheral component. At all events, it is now known that, as in the case of the right-ear advantage, the lag effect is greater for the encoded stops than for the unencoded vowels (Kirstein, 1971; Porter, Shankweiler, and Liberman, 1969); it has also been found that highly encoded stops show a more consistent effect than the relatively less encoded liquids and semi-vowels (Porter, 1971). Also relevant is the finding that synthetic stops that differ only in the second-formant transitions show a lag effect but that the second-formant transitions alone (that is, the chirps) do not (Porter, 1971). Such results support the conclusion that this effect, too, may be specific to the special processing of speech.[5]

In sum, there is now a great deal of evidence to support the assertion that man has ready access to physiological devices that are specialized for the purpose of decoding the speech signal and recovering the phonetic message. Those devices make it possible for the human being to deal with the speech code easily and without conscious awareness of the process or its complexity. The code is thus a natural one.

---

[5] One experimental result appears so far not to fit with that conclusion: syllables that differed in a linguistically irrelevant pitch contour nevertheless gave a lag effect (Darwin, in press).

Resistance to Distortion

Everyone who has ever worked with speech knows that the signal holds up well against various kinds of distortion. In the case of sentences, a great deal of this resistance depends on syntactic and semantic constraints, which are, of course, irrelevant to our concern here. But in the perception of nonsense syllables, too, the message often survives attempts to perturb it. This is due largely to the presence in the signal of several kinds of redundancy. One arises from the phonotactic rules of the language: not all sequences of speech sounds are allowable. That constraint is presumably owing, though only in part, to limitations having to do with the possibilities of co-articulation. In any case, it introduces redundancy and may serve as an error-correcting device. The other kind of redundancy arises from the fact that most phonetic distinctions are cued by more than one acoustic difference. Perception of place of production of the stop consonants, for example, is normally determined by transitions of the second formant, by transitions of the third formant, and by the frequency position of a burst of noise. Each of these cues is more or less sufficient, and they are highly independent of each other. If one is wiped out, the others remain.

There is one other way in which speech resists distortion that may be the most interesting of all because it implies for speech a special biological status. We refer here to the fact that speech remains intelligible even when it is removed about as completely as it can be from its normal, naturalistic context. In the synthetic patterns so much used by us and others, we can, and often do, play fast and loose with the nature of the vocal-tract excitation and with such normally fixed characteristics of the formants as their number, bandwidth, and relative intensity. Such departures from the norm, resulting in the most extreme cases in highly schematic representations, remain intelligible. These patterns are more than mere cartoons, since certain specific cues must be retained. As Mattingly (in this Status Report) has pointed out, speech might be said in this respect to be like the sign stimuli that the ethologist talks about. Quite crude and unnatural models such as Tinbergen's (1951) dummy sticklebacks, elicit responses provided only that the model preserves the significant characters of the original display. As Manning (1969:39) says,"sign stimuli will usually be involved where it is important never to miss making a response to the stimulus." More generally, sign stimuli are often found when the correct transmission of information is crucial for the survival of the individual or the species. Speech may have been used in this way by early man.

How to Tell Speech from Nonspeech

For anyone who uses the speech code, and especially for the very young child who is in the process of acquiring it, it is necessary to distinguish the sounds of speech from other acoustic stimuli. How does he do this? The easy, and probably wrong, answer is that he listens for certain acoustic stigmata that mark the speech signal. One thinks, for example, of the nature of the vocal-tract excitation or of certain general characteristics of the formants. If the listener could identify speech on the basis of such relatively fixed markers, he would presumably decide at a low level of the perceptual system whether a particular signal was speech or not and, on the basis of that decision, send it to the appropriate processors. But we saw in the preceding section that speech remains speech even when the signal is reduced to an extremely schematic form. We suspect, therefore, that the distinction between speech and nonspeech is not made at some early stage on the basis of general acoustic characteristics.

More compelling support for that suspicion is to be found in a recent experiment by T. Rand (pers. comm.) To one ear he presented all of the first formant, including the transitions, together with the steady-state parts of the second and third formants; when presented alone, these patterns sound vaguely like [da]. To the other ear, with proper time relationships carefully preserved, were presented the 50-msec second-formant and third-formant transitions; alone, these sound like the chirps we have referred to before. But when these patterns were presented together--that is, dichotically--listeners clearly heard [ba], [da] or [ga] (depending on the nature of the second-formant and third-formant transitions) in one ear and, simultaneously, nonspeech chirps in the other. Thus, it appears that the same acoustic events--the second-formant or third-formant transitions--can be processed simultaneously as speech and nonspeech. We should suppose, then, that the incoming signal goes indiscriminately to speech and nonspeech processors. If the speech processors succeed in extracting phonetic features, then the signal is speech; if they fail, then the signal is processed only as nonspeech. We wonder if this is a characteristic of all so-called sign stimuli.

Security of the Code

The speech code is available to all members of the human race, but probably to no other species. There is now evidence that animals other than man, including even his nearest primate relatives, do not produce phonetic strings and their encoded acoustic correlates (Lieberman, 1968, 1971; Lieberman, Klatt, and Wilson, 1969; Lieberman, Crelin, and Klatt, in press). This is due, at least in part, to gross differences in vocal-tract anatomy between man and all other animals. (It is clear that speech in man is not simply an overlaid function, carried out by peripheral structures that evolved in connection with other more fundamental biological processes; rather, some important characteristics of the human vocal tract must be supposed to have developed in evolution specifically in connection with speech.) Presumably, animals other than man lack also the mechanisms of neurological control necessary for the organization and coordination of the gestures of speech, but hard evidence for this is lacking. Unfortunately, we know nothing at all about how animals other than man perceive speech. Presumably, they lack the special processor necessary to decode the speech signal. If so, we must suppose that their perception of speech would be different from ours. They should not hear categorically, for instance, and they should not hear the [di]-[du] patterns of Figure 3 as two-segment syllables which have the first segment in common. Thus, we should suppose that animals other than man can neither produce nor correctly perceive the speech code. If all our enemies were animals other than man, cryptanalysts would have nothing to do--or else they might have the excessively difficult task of breaking an animal code for which man has no natural key.

Subcodes

Our discussion so far has, perhaps, left the impression that there is only one speech code. In one sense this is true, for it appears that there is a universal ensemble of phonetic features defined by the communicative possibilities of the vocal tract and the neural speech processor. But the subset of phonetic features which are actually used varies from language to language. Each language thus has its own phonetic "subcode." A given phonetic feature, however, will be articulated and perceived in the same way in every language in which it is used. Thus, we should be very surprised, for instance, to find a language in which the perception of place for stops was not categorical. If, as Eimas's results lead us to suppose, a child is born with an intuitive knowledge of the universal phonetics, part of his task in learning his native language is to identify the features of its phonetic subcode and to forget the others. These unused features cannot be entirely lost, however, since people do learn how to speak and understand more than one language. But there is some evidence that bilinguals listening to their second language do not necessarily use the same speech cues as native speakers of the language do (Haggard, 1971b).

Secondary Codes

A speaker-hearer can become aware of certain aspects of the linguistic process, in particular its phonological and phonetic processes. The awareness can then be exploited to develop "secondary codes," which may be thought of as additional pseudolinguistic rules added to those of the language. A simple example is a children's "secret language," such as Pig Latin, in which a rule for metathesis and insertion applies to each word. We should suppose that to speak or understand Pig Latin fluently would require not only the unconscious knowledge of the linguistic structure of English that all native speakers have, but also a conscious awareness of a particular aspect of this structure--the phonological segmentation--and a considerable amount of practice. There is evidence, indeed, that speakers of English who lack a conscious awareness of phonological segmentation do not master Pig Latin, despite the triviality of its rules (Savin, in press). The pseudolinguistic character of Pig Latin explains why even a speaker of English who does not know Pig Latin would not mistake it for a natural foreign language, and why one continues to feel a sense of artificiality in speaking it long after he has mastered the trick.

Systems of versification are more important kinds of secondary codes. For a literate society the function of verse is primarily esthetic, but for preliterate societies, verse is a means of transmitting verbal information of cultural importance with a minimum of paraphrase. The rules of verse are, in effect, an addition to the phonology which requires that recalled material not only should preserve the semantic values of the original, but should also conform to a specific, rule-determined phonetic pattern. Thus in Latin epic poetry, a line of verse is divided into six feet, each of which must have one of several patterns of long and short syllables. The requirement to conform to this pattern excludes almost all possible renditions other than the correct one and makes memorization easier and recall more accurate. Since versification rules are in general more elaborate than those of Pig Latin, a greater degree of linguistic awareness is necessary to compose verse. This more complex skill has thus traditionally been the specialized occupation of a few members of a society, though a passive form of the skill, permitting the listener to distinguish "correct" from "incorrect" lines without scanning them syllable by syllable, has been possible for a much larger number of people.

Writing, like versification, is also a secondary code for transmitting verbal information accurately, and the two activities have more in common than might at first appear. The reader is given a visually coded representation of the message, and this representation, whether ideographic, syllabic, or alphabetic, provides very incomplete information about the linguistic structure and semantic content of the message. The skilled reader, however, does not need complete information and ordinarily does not even need all of the partial information given by the graphic patterns but rather just enough to exclude most of the other messages which might fit the context. Being competent in his language, knowing the rules of the writing system, and having some degree of linguistic awareness, he can reproduce the writer's message in reasonably faithful fashion. (Since the specific awareness required is awareness of phonological segmentation, it is not surprising that Savin's group of English speakers who cannot learn Pig Latin also have great difficulty in learning to read.)

The reader's reproduction is not, as a rule, verbatim; he makes small deviations which are acceptable paraphrases of the original and overlooks or, better, unconsciously corrects misprints. This suggests that reading is an active process of construction constrained by the partial information on the printed page, just as remembering verse is an active process of construction, constrained, though much less narrowly, by the rules of versification. As Bartlett (1932) noted for the more general case, the processes of perception and recall of verbal material are not essentially different.

For our purposes, the significant fact about pseudolinguistic secondary codes is that, while being less natural than the grammatical codes of language, they are nevertheless far from being wholly unnatural. They are more or less artificial systems based on those aspects of natural linguistic activities which can most readily be brought to consciousness: the levels of phonology and phonetics. All children do not acquire secondary codes maturationally, but every society contains some individuals who, if given the opportunity, can develop sufficient linguistic awareness to learn them, just as every society has its potential dancers, musicians, and mathematicians.

LANGUAGE, SPEECH, AND RESEARCH ON MEMORY

What we have said about the speech code may be relevant to research on memory in two ways: most directly, because work on memory for linguistic information, to which we shall presently turn, naturally includes the speech code as one stage of processing; and, rather indirectly, because the characteristics of the speech code provide an interesting basis for comparison with the kinds of code that students of memory, including the members of this conference, talk about. In this section of the paper we will develop that relevance, summarizing where necessary the appropriate parts of the earlier discussion.

The Speech Code in Memory Research

Acoustic, auditory, and phonetic representations. When a psychologist deals with memory for language, especially when the information is presented as speech sounds, he would do well to distinguish the several different forms that the information can take, even while it remains in the domain of speech. There is, first, the acoustic form in which the signal is transmitted. This is characterized by a poor signal-to-noise ratio and a very high bit rate. The second form, found at an early stage of processing in the nervous system, is auditory. This neural representation of the information maps in a relatively straightforward way onto the acoustic signal. Of course, the acoustic and auditory forms are not identical. In addition to the fact that one is mechanical and the other neural, it is surely true that some information has been lost in the conversion. Moreover, as we pointed out earlier in the paper, it is likely that the signal has been sharpened and clarified in certain ways. If so, we should assume that the task was carried out by devices not unlike the feature detectors the neurophysiologist and psychologist now investigate and that apparently operate in visual perception, as they do in hearing, to increase contrast and extract certain components of the pattern. But we should emphasize that the conversion from acoustic to auditory form, even when done by the kind of device we just assumed, does not decode the signal, however much it may improve it. The relation of the auditory to the acoustic form remains simple, and the bit rate, though conceivably a good deal lower at this neural stage than in the sound itself, is still very high. To arrive at the phonetic representation, the third form that the information takes, requires the specialized decoding processes we talked about earlier in the paper. The result of that decoding is a small number of unitary neural patterns, corresponding to phonetic features, that combine to make the somewhat greater number of patterns that constitute the phonetic segments; arranged in their proper order, these segments become the message conveyed by the speech code. The phonetic representations are, of course, far more economical in terms of bits than the auditory ones. They also appear to have special standing as unitary physiological and biological realities. In general, then, they are well suited for storage in some kind of short-term memory until enough have accumulated to be recoded once more, with what we must suppose is a further gain in economy.

Even when language is presented orthographically to the subjects' eyes, the information seems to be recoded into phonetic form. One of the most recent and also most interesting treatments of this matter is to be found in a paper by Conrad (in press). He concludes, on the basis of considerable evidence, that while it is possible to hold the alphabetic shapes as visual information in short-term memory--deaf-mute children seem to do just that--the information can be stored (and dealt with) more efficiently in phonetic form. We suppose that this is so because the representations of the phonetic segments are quite naturally available in the nervous system in a way, and in a form, that representations of the various alphabetic shapes are not. Given the complexities of the conversion from acoustic or auditory form to phonetic, and the advantages for storage of the phonetic segments, we should insist that this is an important distinction.

Storage and transmission in man and machine. We have emphasized that in spoken language the information must be in one form (acoustic) for transmission and in a very different form (phonetic or semantic) for storage, and that the conversion from the one to the other is a complex recoding. But there is no logical requirement that this be so. If all the components of the language system had been designed from scratch and with the same end in view, the complex speech code might have been unnecessary. Suppose the designer had decided to make do with a smaller number of empty segments, like the phones we have been talking about, that have to be transmitted in rapid succession. The engineer might then have built articulators able to produce such sequences simply--alphabetically or by a cipher--and ears that could perceive them. Or if he had, for some reason, started with sluggish articulators and an ear that could not resolve rapid-fire sequences of discrete acoustic signals, he might have used a larger inventory of segments transmitted at a lower rate. In either case the information would not have had to be restructured in order to make it differentially suitable for transmission and storage; there might have been, at most, a trivial conversion by means of a simple cipher. Indeed, that is very much the situation when computers "talk" to each other. The fact that the human being cannot behave so simply, but must rather use a complex code to convert between transmitted sound and stored message, reflects the conflicting design features of components that presumably developed separately and in connection with different biological functions. As we noted in an earlier part of the paper, certain structures, such as the vocal tract, that evolved originally in connection with nonlinguistic functions have undergone important modifications that are clearly related to speech. But these adaptations apparently go only so far as to make possible the further matching of components brought about by devices such as those that underlie the speech code.

It is obvious enough that the ear involved long before speech made its appearance, so we are not surprised, when we approach the problem from that point of view, to discover that not all of its characteristics are ideally suited to the perception of speech. But when we consider speech production and find that certain design features do not mesh with the characteristics of the ear, we are led to wonder if there are not aspects of the process--in particular, those closer to the semantic and cognitive levels--that had independently reached a high state of evolutionary development before the appearance of language as such and had then to be imposed on the best available components to make a smoothly functioning system. Indeed, Mattingly (this Status Report) has explicitly proposed that language has two sources, an intellect capable of semantic representation and a system of "social releasers" consisting of articulated sounds, and that grammar evolved as an interface between these two very different mechanisms.

In the alphabet, man has invented a transmission vehicle for language far simpler than speech--a secondary code, in the sense discussed earlier. It is a straightforward cipher on the phonological structure, one optical shape for each phonological segment, and has a superb signal-to-noice ratio. We should suppose that it is precisely the kind of transmission vehicle that an engineer might have devised. That alphabetic representations are, indeed, good engineering solutions is shown by the relative ease with which engineers have been able to build the so-called optical character readers. However, the simple arrangements that are so easy for machines can be hard for human beings. Reading comes late in the child's development; it must be taught; and many fail to learn. Speech, on the other hand, bears a complex relation to language as we have seen and has so far defeated the best efforts of engineers to build a device that will perceive it. Yet this complex code is mastered by children at an early age, some significant proficiency being present at four weeks; it requires no tuition; and everyone who can hear manages to perceive speech quite well.

The relevance of all this to the psychology of memory is an obvious and generally observed caution: namely, that we be careful about explaining human beings in terms of processes and concepts that work well in intelligent and remembering machines. We nevertheless make the point because we have in speech a telling object lesson. The speech code is an extremely complex contrivance, apparently designed to make the best of a bad fit between the requirement that phonetic segments be transmitted at a rapid rate and the inability of the mouth and the ear to meet that requirement in any simple way. Yet the physiological devices that correct this mismatch are so much a part of our being that speech works more easily and naturally for human beings than any other arrangement, including those that are clearly simpler.

More and less encoded elements of speech. In describing the characteristics of the speech code we several times pointed to differences between stop consonants and vowels. The basic difference has to do with the relation between signal and message: stop consonants are always highly encoded in production, so their perception requires a decoding process; vowels can be, and sometimes are, represented by encipherment, as it were alphabetically, in the speech signal, so they might be perceived in a different and simpler way. We are not surprised, then, that stops and vowels differ in their tendencies toward categorical perception as they do also in the magnitude of the right-ear advantage and the lag effect (see above).

An implication of this characteristic of the speech code for research in immediate memory has appeared in a study by Crowder (in press) which suggests that vowels produce a "recency" effect, but stops do not. Crowder and Morton (1969) had found that, if a list of spoken words is presented to a subject, there is an improvement in recall for the last few items on the list, but no such recency effect is found if the list is presented visually. To explain this model difference, Crowder and Morton suggested that the spoken items are held for several seconds in an "echoic" register in "precategorical" or raw sensory form. At the time of recall these items are still available to the subject in all their original sensory richness and are therefore easily remembered. When presented visually, the items are held in an "iconic" store for only a fraction of a second. In his more recent experiment Crowder has found that for lists of stop-vowel syllables, the auditory recency effect appears if the syllables on the list contrast only in their vowels but is absent if they contrast only in their stops. If Crowder and Morton's interpretation of their 1969 result is correct, at least in general terms, then the difference in recency effect between stops and vowels is exactly what we should expect. As we have seen in this paper, the special process that decodes the stops strips away all auditory information and presents to immediate perception a categorical linguistic event the listener can be aware of only as [b,d,g,p,t, or k]. Thus, there is for these segments no auditory, precategorical form that is available to consciousness for a time long enough to produce a recency effect. The relatively unencoded vowels, on the other hand, are capable of being perceived in a different way. Perception is more nearly continuous than categorical: the listener can make relatively fine discriminations within phonetic classes because the auditory characteristics of the signal can be preserved for a while. (For a relevant model and supporting data see Fujisaki and Kawashima, 1969.) In the experiment by Crowder, we may suppose that these same auditory characteristics of the vowel, held for several seconds in an echoic sensory register, provide the subject with the rich, precategorical information that enables him to recall the most recently presented items with relative ease.

It is characteristic of the speech code, and indeed of language in general, that not all elements are psychologically and physiologically equivalent. Some (e.g., the stops) are more deeply linguistic than others (e.g., the vowels); they require special processing and can be expected to behave in different ways when memory codes are used.

Speech as a special process. Much of what we said about the speech code was to show that it is complex in a special way and that it is normally processed by a correspondingly special device. When we examine the formal aspects of this code, we see resemblances of various kinds to the other grammatical codes of phonology and syntax--which is to say that speech is an integral part of a larger system called language--but we do not readily find parallels in other kinds of perception. We know very little about how the speech processor works, so we cannot compare it very directly with other kinds of processors that the human being presumably uses. But knowing that the task it must do appears to be different in important ways from the tasks that confront other processors, and knowing, too, that the speech processor is in one part of the brain while nonspeech processors are in another, we should assume that speech processing may be different from other kinds. We might suppose, therefore, that the mechanisms underlying memory for linguistic information may be different from those used in other kinds of memory such as, for example, visual or spatial.

Speech appears to be specialized, not only by comparison with other perceptual or cognitive systems of the human being, but also by comparison with any of the systems so far found in other animals. While there may be some question about just how many of the so-called higher cognitive and linguistic processes monkeys are capable of, it seems beyond dispute that the speech code is unique to man. To the extent, then, that this code is used in memory processes--for example, in short-term memory--we must be careful about generalizing results across species.

Speech and Memory Codes Compared

It will be recalled that we began by adopting the view that paraphrase has more to do with the processes by which we remember than with those by which we forget. In this vein we proposed that when people are presented with long stretches of sensible language, they normally use the devices of grammar to recode the information from the form in which it was transmitted into a form suitable for storage. On the occasion of recall they code it back into another transmittable form that may resemble the input only in meaning. Thus, grammar becomes an essential part of normal memory processes and of the memory codes that this conference is about. We therefore directed our attention to grammatical codes, taking these to be the rules by which conversions are carried out from one linguistic level to another. To spell out the essential features of such codes, we chose to deal in detail with just one, the speech code. It can be argued, persuasively we think, that the speech code is similiar to other grammatical codes, so its characteristics can be used, within reasonable limits, to represent those of grammar generally. But speech has the advantage in this connection that it has been more accessible to psychological investigation than the other grammatical codes. As a result, there are experimental data that permit us to characterize speech in ways that provide a useful basis for comparison with the codes that have come from the more conventional research on verbal memory. In this final section we turn our attention briefly to those more conventional memory codes and to a comparison between them and the speech code.

We will apply the same convention to this discussion of conventional memory codes that we applied to our discussion of grammatical codes. That is, the term "code" is reserved for the rules which convert from one representation of the information to another. In our analysis of the speech code we took the acoustic and phonetic levels as our two representations and inferred the properties of the speech code from the relation between the two.

In the most familiar type of experiment the materials the subject is required to remember are not the longer segments of language, such as sentences or discourses, but rather lists of words or nonsense syllables. Typically in such an experiment, the subject is required to reproduce the information exactly as it was presented to him, and his response is counted as an error if he does not. Under those circumstances it is difficult, if not impossible, for the subject to employ his linguistic coding devices to their fullest extent, or in their most normal way. However, it is quite evident that the subject in this situation nevertheless uses codes; moreover, he uses them for the same general purpose to which, we have argued, language is so often put, which is to enable him to store the information in a form different from that in which it was presented. Given the task of remembering unfamiliar sequences such as consonant trigraphs, the subject may employ, sometimes to the experimenter's chagrin, some form of linguistic mediation (Montague, Adams, and Kiess, 1966). That is, he converts the consonant sequence into a sentence or proposition, which he then stores along with a rule for future recovery of the consonant string. In a recent examination of how people remember nonsense syllables, Prytulak (1971) concluded that such mediation is the rule rather than the exception. Reviewing the literature on memory for verbal materials, Tulving and Madigan (1970) describe two kinds of conversions: one is the substitution of an alternative symbol for the input stimulus together with a conversion rule; the other is the storage of ancillary information along with the to-be-remembered item. Most generally, it appears that when a subject is required to remember exactly lists of unrelated words, paired-associates, or digit strings, he tries to impart pattern to the material, to restructure it in terms of familiar relationships. Or he resorts, at least in some situations, to the kind of "chunking" that Miller (1956) first described and that has become a staple of memory theory (Mandler, 1967). Or he converts the verbal items into visual images (Paivio, 1969; Bower, 1970). At all events, we find that, as Bower (1970) has pointed out, bare-bones rote memorization is tried only as a last resort, if at all.

The subject converts to-be-remembered material which is unrelated and relatively meaningless into an interconnected, meaningful sequence of verbal items or images for storage. What can be said about the rules relating the two levels? In particular, how do the conversions between the two levels compare with those that occur in the speech code, and thus, indirectly, in language in general? The differences would appear to be greater than the similarities. Many of these conversions that we have cited are more properly described as simple ciphers than as codes, in the sense that we have used these terms earlier, since there is in these cases no restructuring of the information but only a rather straightforward substitution of one representation for another. Moreover, memory codes of this type are arbitrary and idiosyncratic, the connection between the two forms of the information having arisen often out of the accidents of the subject's life history; such rules as there may be (for example, to convert each letter of the consonant trigraph to a word beginning with that letter) do not truly rationalize the code but rather fall back, in the end, on a key that is, in effect, a code book. As often as not, the memory codes are also relatively unnatural: they require conscious effort and, on occasion, are felt by the subject to be difficult and demanding. In regard to efficiency, it is hard to make a comparison; relatively arbitrary and unnatural codes can nevertheless be highly efficient given enough practice and the right combination of skills in the user.

In memory experiments which permit the kind of remembering characterized by paraphrase, we would expect to find that memory codes would be much like language codes, and we should expect them to have characteristics similar to those of the code we know as speech. The conversions would be complex recodings, not simple substitutions; they would be capable of being rationalized; and they would, of course, be highly efficient for the uses to which they were being put. But we would probably find their most obvious characteristic to be that of naturalness. People do not ordinarily contrive mnemonic aids by which to remember the gist of conversations or of books, nor do they necessarily devise elaborate schemes for recalling stories and the like, yet they are reasonably adept at such things. They remember without making an effort to commit a message to memory; more important, they do not have to be taught how to do this sort of remembering.

It is, of course, exceedingly difficult to do scientific work in situations that permit the free use of these very natural language codes. Proper controls and measures are hard to arrange. Worse yet, the kinds of paraphrase that inevitably occur in long discourses will span many sentences and imply recoding processes so complex that we hardly know now how to talk about them. Yet, if the arbitrary, idiosyncratic ciphers which we have described are simply devices to mold to-be-remembered, unrelated materials into a form amenable to the natural codes, then it must be argued that our understanding of such ciphers will advance more surely with knowledge of the natural bases from which they derive and to which they must, presumably, be anchored.

REFERENCES

Bartlett, F.C. (1932) Remembering. (Cambridge, England: Cambridge University Press).

Bower, G.H. (1970) Organizational factors in memory. Cog. Psychol. 1, 18-46.

Broadbent, D.E. and Gregory, M. (1964) Accuracy of recognition for speech presented to the right and left ears. Quart. J. exp. Psychol. 16, 359-360.

Bryden, M.P. (1963) Ear preference in auditory perception. J. exp. Psychol. 65, 103-105.

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: M.I.T. Press).

Chomsky, N. and Halle, M. (1968) The Sound Pattern of English. (New York: Harper and Row).

Conrad, R. (in press) Speech and reading. In *Language by Ear and by Eye: The Relationships between Speech and Reading*, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

Cooper, F.S. (1966) Describing the speech process in motor command terms. J. acoust. Soc. Amer. *39*, 1221A. (Text in Haskins Laboratories Status Report on Speech Research *SR-5/6*, 1966.)

Crowder, R. (in press) The sound of vowels and consonants in immediate memory. J. verb. Learn. verb Behav., *10*.

Crowder, R.B. and Morton, J. (1969) Precategorical and acoustic storage (PAS). Perception and Psychophysics *5*, 365-373.

Darwin, C.J. (1969) *Auditory Perception and Cerebral Dominance*. Unpublished doctoral dissertation, University of Cambridge.

Darwin, C.J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. exp. Psychol. *23*, 46-62.

Darwin, C.J. (in press) Dichotic backward masking of complex sounds. Quart. J. exp. Psychol.

Dorman, M. (1971) *Auditory Evoked Potential Correlates of Speech Perception*. Unpublished doctoral dissertation, University of Connecticut.

Eimas, P.D. (1963) The relation between identification and discrimination along speech and nonspeech continua. Language and Speech *3*, 206-217.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., and Vigorito, J. (1971) Speech perception in infants. Science *171*, 303-306.

Fry, D.B., Abramson, A.S., Eimas, P.D. and Liberman, A.M. (1962) The identification and discrimination of synthetic vowels. Language and Speech *5*, 171-189.

Fujisaki, H. and Kawashima, T. (1969) On the modes and mechanisms of speech perception. In *Annual Report No. 1*. (Tokyo: University of Tokyo, Division of Electrical Engineering, Engineering Research Institute).

Haggard, M.P. (1970) Theoretical issues in speech perception. In *Speech Synthesis and Perception 4*. (Cambridge, England: Psychological Laboratory).

Haggard, M.P. (1971a) Encoding and the REA for speech signals. Quart. J. exp. Psychol. *23*, 34-45.

Haggard, M.P. (1971b) New demonstrations of categorical perception. In *Speech Synthesis and Perception 5*. (Cambridge, England: Psychological Laboratory).

Haggard, M.P., Ambler, S. and Callow, M. (1969) Pitch as a voicing cue. J. acoust. Soc. Amer. *47*, 613-617.

Haggard, M.P. and Parkinson, A.M. (1971) Stimulus and task factors as determinants of ear advantages. Quart. J. exp. Psychol. *23*, 168-177.

Halwes, T. (1969) *Effects of Dichotic Fusion on the Perception of Speech*. Unpublished doctoral dissertation, University of Minnesota.(Reproduced as Supplement to Haskins Laboratories Status Report on Speech Research 1969.)

Kimura, D. (1961) Cerebral dominance and perception of verbal stimuli. Canad. J. Psychol. *15*, 166-171.

Kimura D. (1964) Left-right differences in the perception of melodies. Quart. J. exp. Psychol. *16*, 355-358.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex *3*, 163-178.

Kirstein, E. (1971) *Temporal Factors in the Perception of Dichotically Presented Stop Consonants and Vowels*. Unpublished doctoral dissertation, University of Connecticut. (Reproduced in Haskins Laboratories Status Report on Speech Research *SR-24*.)

Kirstein, E. and Shankweiler, D.P. (1969) Selective listening for dichotically presented consonants and vowels. Paper read before 40th Annual Meeting of Eastern Psychological Association, Philadelphia, 1969. (Text in Haskins Laboratories Status Report on Speech Research SR-17/18, 133-141.)

Liberman, A.M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. acoust. Soc. Amer. 44, 1574-1584.

Lieberman, P. (1971) On the speech of Neanderthal man. Linguistic Inquiry 2, 203-222.

Lieberman, P., Klatt, D., and Wilson, W.A. (1969) Vocal tract limitations on the vowel repertoires of rhesus monkeys and other nonhuman primates. Science 164, 1185-1187.

Lieberman, P., Crelin, E.S., and Klatt, D.H. (in press) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. American Anthropologist. (Also in Haskins Laboratories Status Report on Speech Research SR-24, 51-90.)

Lindblom, B. (1963) Spectrographic study of vowel reduction. J. acoust. Soc. Amer. 35, 1773-1781.

Lisker, L. and Abramson, A.S. (1967) Some effects of context on voice onset time in English stops. Language and Speech 10, 1-28.

Mandler, G. (1967) Organization and memory. In The Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 1, K.W. Spence and J.T. Spence, eds. (New York: Academic Press).

Manning, A. (1969) An Introduction to Animal Behavior. (Reading, Mass.: Addison-Wesley).

Mattingly, I.G. (This Status Report) Speech cues and sign stimuli.

Mattingly, I.G. and Liberman, A.M. (1969) The speech code and the physiology of language. In Information Processing in the Nervous System. K.N. Leibovic, ed. (New York: Springer Verlag).

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., and Halwes, T. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Miller, G.A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63, 81-97.

Montague, W.E., Adams, J.A., and Kiess, H.O. (1966) Forgetting and natural language mediation. J. exp. Psychol. 72, 829-833.

Ohman, S.E.G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. acoust. Soc. Amer. 39, 151-168.

Paivio, A. (1969) Mental imagery in associative learning and memory. Psychol. Rev. 76, 241-263.

Pisoni, D. (1971) On the Nature of Categorical Perception of Speech Sounds. Unpublished doctoral dissertation, University of Michigan. (Reproduced as Supplement to Haskins Laboratories Status Report on Speech Research, 1971.)

Porter, R.J. (1971) Effects of a Delayed Channel on the Perception of Dichotically Presented Speech and Nonspeech Sounds. Unpublished doctoral dissertation, University of Connecticut.

Porter, R., Shankweiler, D.P., and Liberman, A.M. (1969) Differential effects of binaural time differences in perception of stop consonants and vowels. Paper presented at annual meeting of the American Psychological Association, Washington, D.C., 2 September.

Prytulak, L.S. (1971) Natural language mediation. Cog. Psychol. 2, 1-56.

Savin, H. (in press) What the child knows about speech when he starts learning to read. In Language by Ear and by Eye: The Relationships between Speech and Reading, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press).

Shankweiler, D. and Studdert-Kennedy, M. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. exp. Psychol. 19, 59-63.

Spellacy, F. and Blumstein, S. (1970) The influence of language set on ear preference in phoneme recognition. Cortex 6, 430-439.

Stevens, K.N., Liberman, A.M., Ohman, S.E.G., and Studdert-Kennedy, M. (1969) Cross-language study of vowel perception. Language and Speech 12, 1-23.

Studdert-Kennedy, M. (in press) The perception of speech. In Current Trends in Linguistics, T.A. Sebeok, ed. (The Hague: Mouton). (Also in Haskins Laboratories Status Report on Speech Research SR-23, 15-48.)

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., and Cooper, F.S. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Studdert-Kennedy, M. and Shankweiler, D. (1970) Hemispheric specialization for speech perception. J. acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., Shankweiler, D., and Schulman, S. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. acoust. Soc. Amer. 48, 599-602.

Tinbergen, N. (1951) The Study of Instinct. (Oxford: Clarendon Press).

Tulving, E. and Madigan, S.A. (1970) Memory and verbal learning. Annual Rev. Psychol. 21, 437-484.

Vinegrad, M. (1970) A direct magnitude scaling method to investigate categorical versus continuous modes of speech perception. Haskins Laboratories Status Report on Speech Research SR-21/22, 147-156.

Warren, R.M., Obusek, C.J., Farmer, R.M., and Warren, R.T. (1969) Auditory sequence: Confusions of patterns other than speech or music. Science 164, 586-587.

Speech Cues and Sign Stimuli*

Ignatius G. Mattingly+
Haskins Laboratories, New Haven

The perception of the linguistic information in speech, as investigations carried on over the past twenty years have made clear, depends not on a general resemblance between presently and previously heard sounds but on a quite complex system of acoustic cues which has been called by Liberman et al. (1967) the "speech code." These authors suggest that a special perceptual mechanism is used to detect and decode the speech cues. I wish to draw attention here to some interesting formal parallels between these cues and a well-known class of animal signals, "sign stimuli," described by Lorenz, Tinbergen, and others. These formal parallels suggest some speculations about the original biological function of speech and the related problem of the origin of language.

A speech cue is a specific event in the acoustic stream of speech which is important for the perception of a phonetic distinction. A well-known example is the second-formant transition, a cue to place of articulation. During speech, the formants (i.e., acoustical resonances) of the vocal tract vary in frequency from moment to moment depending on the shape and size of the tract (Fant, 1960). When the tract is excited (either by periodic glottal pulsing or by noise) these momentary variations can be observed in a sound spectrogram. During the transition from a stop consonant, such as [b,d,g,p,k], to a following vowel, the second (next to lowest in frequency) formant (F2) moves from a frequency appropriate for the stop towards a frequency appropriate for the vowel; the values of these frequencies depend mainly on the position of the major constriction of the vocal tract in the formation of each of the two sounds. Since there is no energy in most or all of the acoustic spectrum until after the release of the stop closure, the earlier part of the transition will be neither audible nor observable. But the slope of the later part, following the release, is audible and can be observed (see the transition for [b] in the spectrogram for [bɛ] in the upper portion of Figure 1). It is also a sufficient cue to the place of articulation of the preceding stop: labial [b,p], alveolar [d,t], or velar [g,k]. It is as if the listener, given the final part of the F2 transition, could extrapolate back to the consonantal frequency or locus (Delattre et al., 1955).

---

*Paper to appear in American Scientist (1972) in press.

+Also University of Connecticut, Storrs.

Acknowledgments: The preparation of this paper was supported in part by the Fulbright-Hays Commission and the Provost and Fellows of King's College, Cambridge. I also wish to acknowledge gratefully encouragement and criticism given by G. Evelyn Hutchinson, Alvin M. Liberman, Benjamin Sachs, Jacqueline Sachs, Michael Studdert-Kennedy, Philip Lieberman, Alison Jolly, Mark Haggard, Adrian Fourcin, and Dennis Fry, but responsibility for errors is mine.

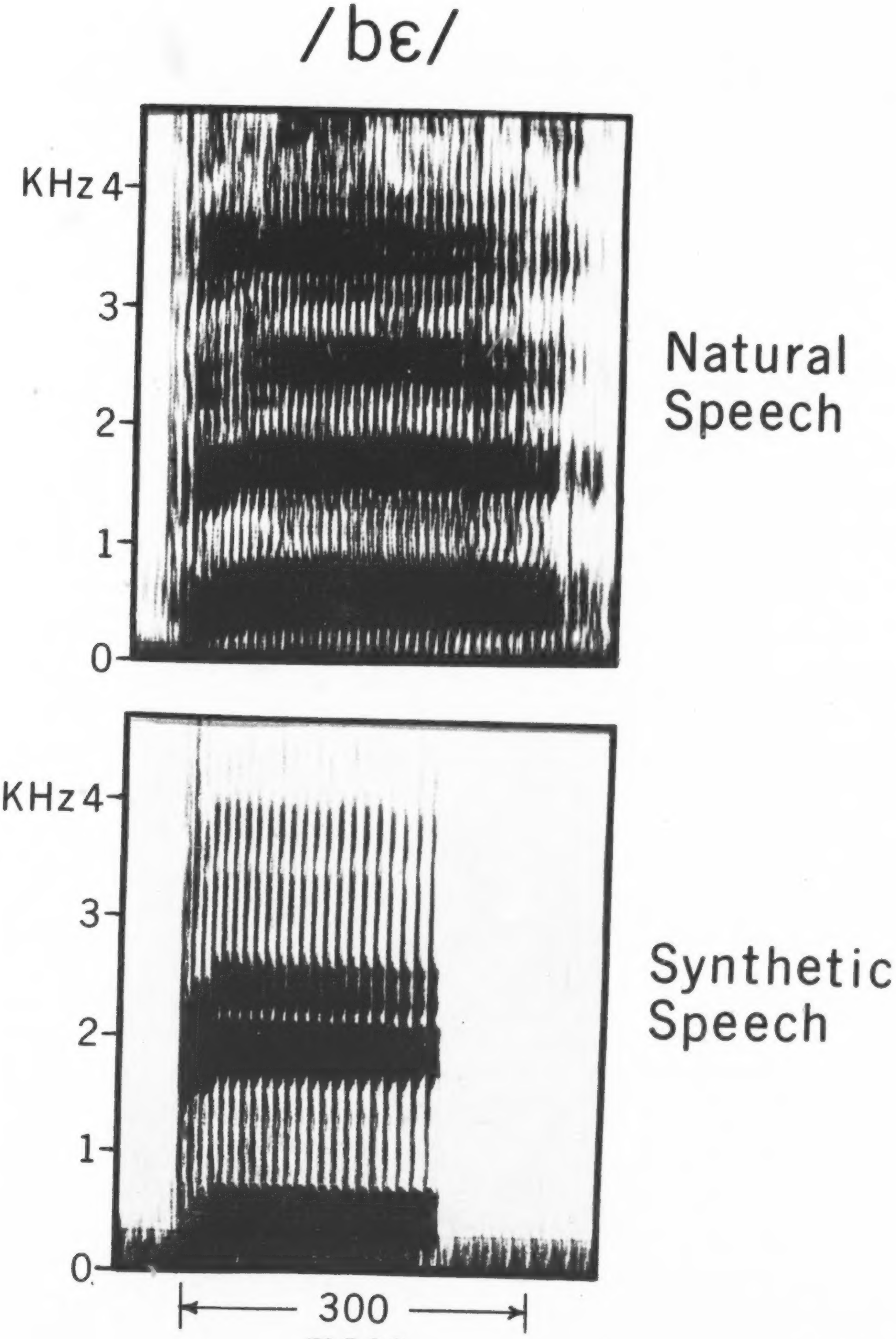


Fig. I

Spectrograms of Natural and Synthetic Speech for [bɛ]

It is possible electronically to synthesize speech which is intelligible, even though it has much simpler spectral structure than natural speech (Cooper, 1950; Mattingly, 1968). In the lower portion of Figure 1 is shown a spectrogram of a synthetic version of the syllable [bɛ]. Synthetic speech can be used to demonstrate the value of a cue such as the F2 transition by generating a series of stop-vowel syllables for which the slope of the audible part of the F2 transition is the only variable, and other cues to position of articulation, such as the frequency of the burst of noise following the release of the stop, or the slope of F3, are absent or neutralized (Cooper et al., 1952). A syllable in a series such as this will be heard as beginning with a labial, an alveolar, or a velar stop depending entirely on the slope of the F2 transition. This is true even though the slope values appropriate for a particular stop consonant depend on the vowel: thus a rising F2 cues [d] before [i], and a falling F2, [d] before [u] (see the patterns in Figure 3).

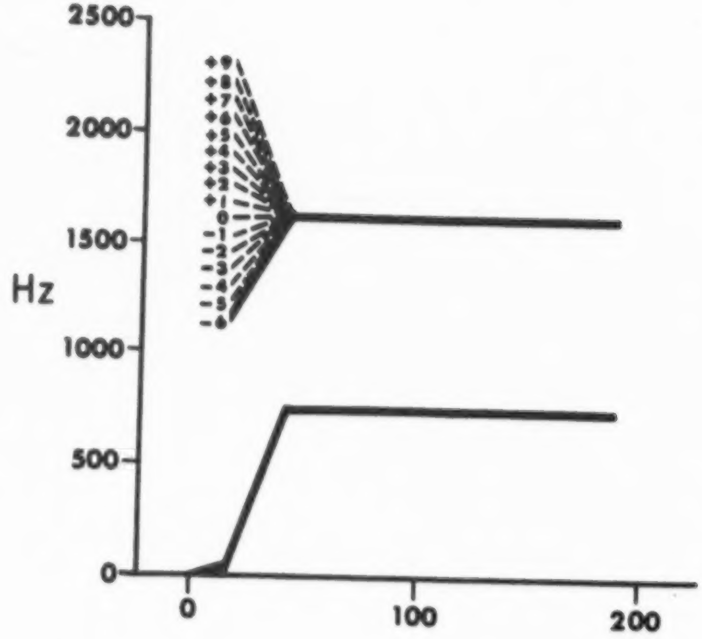
Phonetic distinctions other than place are signalled by other cues. Thus, in English, the cue separating the voiceless, aspirated stops [p,t,k] from the voiced stops [b,d,g] is voice-onset time (Liberman et al., 1958). If the beginning of glottal pulsing coincides with, or precedes, the release, the stop will be heard as [b], [d], or [g], depending upon the cues to place of articulation; if the pulsing is delayed 30 msec or more after the release, the stop will be heard as [p], [t], or [k]. Again, the duration of the formant transitions is a cue for the stop-semivowel distinction (e.g., [b] vs. [w]) (Liberman et al., 1956). A shorter (30-40 msec) transition will be heard as a stop, whereas a longer (60-80 msec) transition will be heard as a semivowel.

Some recent work indicates that human beings may possibly be born with knowledge of these cues. While appropriate investigations have not yet been carried out for most of the cues, the facts with respect to voice-onset time are rather suggestive. Not all languages have this distinction between stops with immediate voice onset and stops with voice onset delayed after release, but for all those that do, the amount of delay required for a stop to be heard as voiceless rather than voiced is about the same (Lisker and Abramson, 1970; Abramson and Lisker, 1970). This constraint on perception thus appears to be a true language universal, and so likely to reflect a physiological limitation rather than a learned convention.

Exploring the question more directly, Eimas et al. (1970), by monitoring changes in the sucking rate of one-month-old infants listening to synthetic speech stimuli, showed that the infants could distinguish significantly better between two stop-vowel stimuli which straddle the critical value of voice-onset time than between two stimuli which do not, even though the absolute difference in voice-onset time is the same. Thus the information required to interpret at least one speech cue appears either to be learned with incredible speed or to be genetically transmitted.

Sign stimuli, with which I propose to compare speech cues, have been defined by Russell (1943), Tinbergen (1951), and other ethologists as simple, conspicuous, and specific characters of a display which under given conditions produces an "instinctive" response: the red belly of the male stickleback, which provokes a rival to attack, or the zigzag pattern of his dance, which

Patterns for a Series of Stop-Vowel Syllables with Systematically Varied F2 Transitions



Note: F2 transitions with low starting points will cause the stop to be heard as [b], those with high starting points as [g], and those in between as [d].

Fig. 2

F1 and F2 Patterns Heard as [di] and [du], Despite the Apparent Difference in the F2 Transition

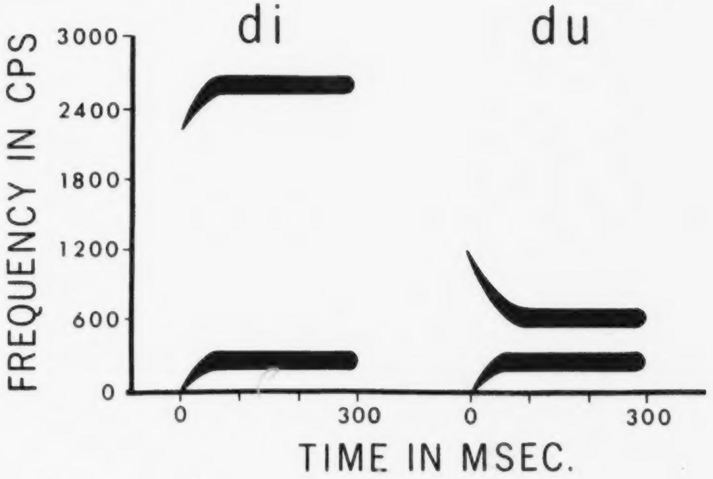


Fig. 3

arouses the female (Tinbergen, 1951); the spots by which the ringed plover identifies her eggs (Koehler and Zagarus, 1937); the red spot on the herring gull's bill, which makes her chicks beg for food (Tinbergen, 1951). These examples are visual, but sign stimuli are found in other modalities also: e.g., the monotone note of the white-throated sparrow's song, by which he asserts his territorial claims (Falls, 1969); or the chemical in the blood from a wounded minnow, which causes other minnows to flee when they scent it in the water (Manning, 1967). Responding properly to sign stimuli is normally of great value for the survival of the individual or the species. As Manning (1967:39) comments, "Sign stimuli will usually be involved where it is important never to miss making a response to the stimulus." It is this circumstance, perhaps, which accounts for the striking properties of sign stimulus perception which we shall be mainly concerned with here: the animal responds not to the display in general but specifically to the sign stimuli, and the strength of the response is in proportion to the number and conspicuousness of the sign stimuli. The perception of a sign stimulus and the response it produces have been attributed by Lorenz (1935) to a special neural "innate releasing mechanism."

The concepts of the sign stimulus and the innate releasing mechanism, as used in early ethological work, have come in for much justified criticism (e.g., Hailman, 1969; Hinde, 1970). It has been argued that sign stimuli cannot be shown to differ in principle from other stimuli; that some purported sign stimuli are not actually specific to particular responses but merely reflect the general capabilities of the animal's sense organs or associated perceptual equipment; that the word "innate" suggests too simple a dichotomy between nature and nurture; and that sign stimuli do not always lead to direct and immediate responses but influence behavior in other ways.

But when all these criticisms are taken into account, there remain some very striking phenomena. There are many cases in which a stimulus is selectively perceived by a particular species and not by others. The selectivity cannot be accounted for simply by an appeal to the general sensory capabilities of the species. The stimulus consistently elicits a direct response (or other specific behavior indicating that the stimulus has been perceived, as in the case of orientation). This response is adaptive. Moreover, in many instances (and in all the examples given above) the stimulus is a character of a display by a conspecific (or symbiotically related) individual; the entire pattern of behavior, consisting of the display and the response, is adaptive.

Displays of this latter sort have been called "social releasers" (Tinbergen, 1951:171). Their component sign stimuli elicit appropriate responses from conspecific individuals in situations important for group safety or for the integrity and continuity of the species. Social releasers include: alarm calls; the "threat behavior" of many species, by which the adaptive ends of sexual fighting are achieved with few actual casualties; the displays which serve as reproductive isolating mechanisms, encouraging intraspecific and discouraging interspecific mating; and the signs by which parents and young identify each other, so that the latter are protected and fed. In all these adaptively important situations, displays composed of sign stimuli serve to authenticate the conspecificity of individuals.

It has also been suggested before that sign stimuli actually occur in human behavior. The facial characteristics and limb movements of babies evoke parental behavior (Tinbergen, 1951). Babies, in turn, respond to adult facial characteristics, notably to eyes and to smiles, and women have a universal flirting gesture (Eibl-Eibesfeldt, 1970). I think that speech cues may also belong to the class of human sign stimuli, despite obvious differences to be discussed shortly. But let us now consider the resemblances.

First of all, the speech cues, like the sign stimuli, do not require a natural context, or even a naturalistic one; the appropriate response can be elicited by drastically simplified models of the natural original. Tinbergen's sticklebacks would respond to an extremely crude model, provided only that it had a red belly, but disdained very naturalistic models which lacked this crucial feature (Figure 4) (Tinbergen, 1951:28). Lorenz (1954: 291, translated by Eibl-Eibesfeldt, 1970:88) makes the general claim that "where an animal can be 'tricked' into responding to simple models, we have a response by an innate releasing mechanism." In the case of speech, most of the complexity of the spectrum can be dispensed with so long as the essential cues are preserved. It has already been mentioned that the simple, two-formant synthetic utterances of Figure 2 are clearly heard by subjects as [b], [d], etc. The natural and synthetic utterances in Figure 1 are linguistically equivalent, even though in the latter only the lower formants appear, and these in a very stylized configuration.

The synthetic utterance is not, however, simply an acoustic cartoon of the natural utterance. Though it shares with a cartoon the appearance of extreme simplicity and emphasis of salient features, it is rather a systematic attempt to represent, consistently but exclusively, the essential acoustic cues, all other details of the signal being discarded or neutralized. The principal loss in such synthetic speech is not intelligibility but only naturalness. This is rather surprising. One might quite reasonably expect that intelligibility would depend crucially on naturalness, that tampering with the observed spectrum of a natural utterance to any degree would alter its linguistic value or cause it not to be perceived linguistically at all. I do not mean to imply that high-quality natural speech would not be more intelligible than synthetic speech, or that sticklebacks would not respond more strongly to a real stickleback with a red belly than to a dummy. In synthetic speech, a host of redundant minor cues, as yet unidentified, are no doubt sacrificed together with the linguistically irrelevant details of the signal. Similarly, in the construction of the dummy, sign stimuli of minor importance have been ignored. But it appears that the dependence of artificial speech cues and sign stimuli on a naturalistic context is very small. Though the listener and (for all we know) the stickleback may be quite aware of the lack of naturalness, neither one appears to be disturbed by it. The relative naturalness of the speech cues and sign stimuli themselves is something else again, as will be seen shortly.

Both speech cues and sign stimuli exhibit what Tinbergen (1951:81), translating Seitz (1940), calls "the phenomenon of heterogeneous summation." The same response can be elicited by separate and noninteracting sign stimuli: thus, either the redness of the patch on the herring gull's bill or the contrast of the patch with the rest of the bill release the chick's pecking response. Moreover, if two stimuli for the same response are present, but one

Stickleback Models Used by Tinbergen

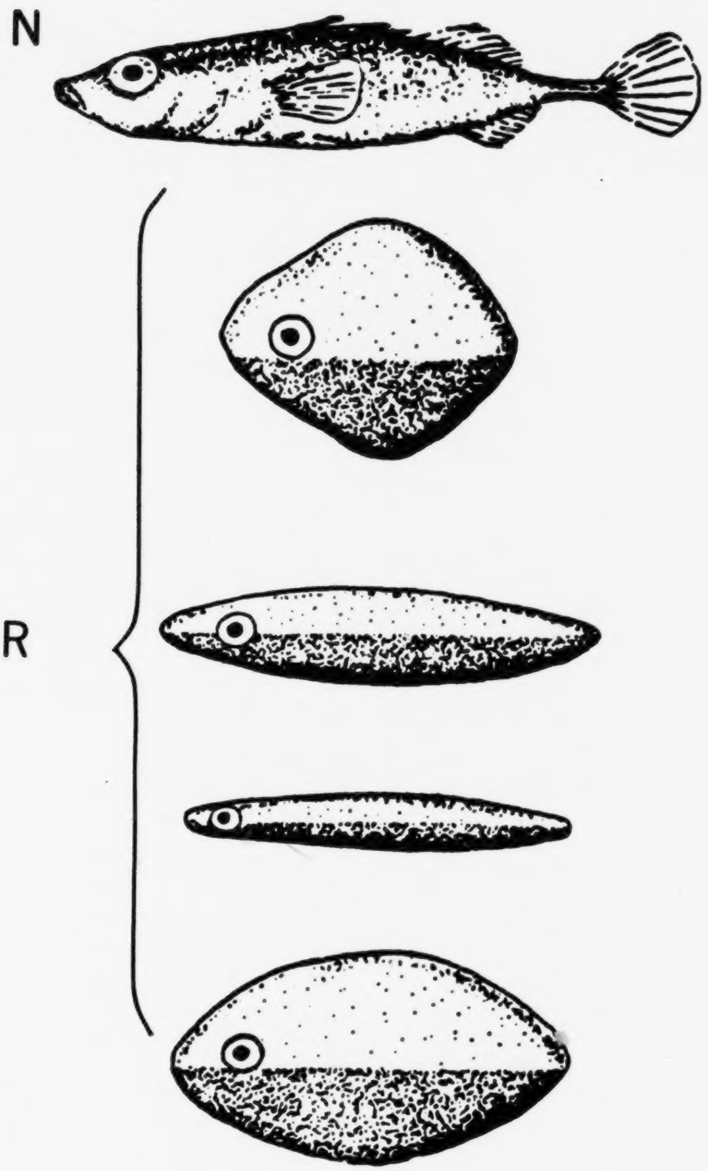


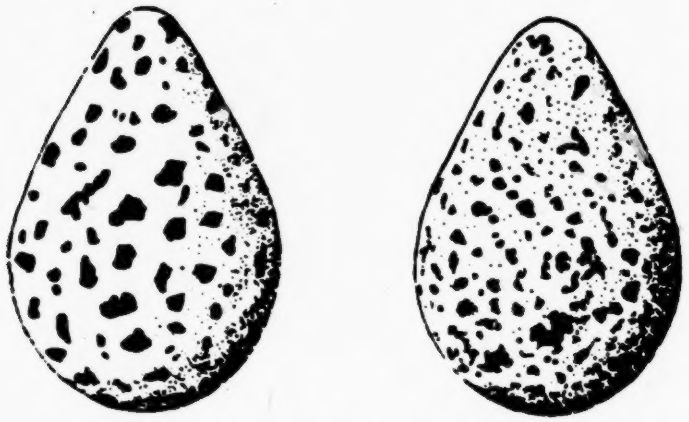
Fig. 4

Note: The fairly realistic model marked N, which lacked a red belly, provoked attack by male sticklebacks much less than the various crude models labeled R, which have red bellies. (After Tingergen, 1951.)

is defective, the second will compensate for the deficiency of the first. A similar principle operates in speech perception. Multiple cues for the same phonetic feature are the rule. For example, point of articulation in stop consonants is cued not only by the F2 transition but also by the F3 transition and by a burst of noise at an appropriate frequency just after release of stop closure (Delattre et al., 1955; Halle et al., 1957; Harris et al., 1958). In medial position, a voiced rather than a voiceless stop is cued by low-frequency periodic energy during closure, by lesser duration of closure, and by greater length of the preceding vowel (Lisker, 1957). Furthermore, the perceptual weight of one cue appears to be independent of that of the others; all combine additively to carry a single phonetic distinction; if a cue is defective or absent, as is very often the case in natural speech, the deficiency is compensated for by the presence of other cues. Thus Hoffman (1958) compared perception of point of articulation for (a) synthetic stop-vowel syllables in which all three cues (burst, F2 transition, F3 transition) were present, (b) syllables in which the burst cue was absent, (c) syllables in which the third formant with its transition was absent, and (d) syllables in which both third formant and burst were absent and only the F2 transition was present. He found that the optimal version of a cue for a particular point of articulation is the same whether presented separately or in combination with other cues; that labeling is most consistent when all three cues are optimal for the same point of articulation; and that an optimal F3 transition would compensate for a nonoptimal burst cue, and conversely. A.M. Liberman (personal communication) points out that speech also carries multiple cues to the sex of the speaker: men's voices differ from women's both in pitch range and in formant frequency range. Thus, neither the perception of speech cues nor that of sign stimuli is a Gestalt (Hinde, 1970).

An optimal speech cue is often not a realistic one; such a cue is the analog of a "supernormal" sign stimulus, such as the pattern of black spots on a white background on the artificial egg (see Figure 5) which the plover prefers to a natural egg with dark brown spots on a light brown background (Koehler and Zagarus, 1937). "The natural situation," Tinbergen (1951:44) observes, "is not always optimal." Similarly, if a human subject is presented with stimuli like those represented in Figure 2, he will hear the first few, those with rising transitions, as [bɛ]. The stimuli with the less steeply sloping transitions are closer to what one observes in instances of [bɛ] in natural speech, while the more extreme transitions are unlikely, perhaps even articulatorily impossible. Yet, in a labeling test, the more steeply rising the F2 transition, the more likely is the subject to hear [bɛ]. Thus the subject will label more consistently not only when more cues are present but also when the cues present are more nearly optimal, i.e., supernormal. Again, vowels spoken in isolation will occupy more extreme positions on the F1-F2 plane than vowels in connected speech (Shearme and Holmes, 1962) and are easier to label than the "same" vowels excised from connected speech. As Manning (1967) says, the failure of a sign stimulus to evolve to the supernormal extreme can usually be explained by considering other functional requirements. Thus the low-contrast, brown-on-brown spotting of the plover's eggs also serves to camouflage them from predators; black on a white background would not be so effective. The vocal tract, likewise, is primarily a group of devices for breathing and eating. A vocal tract which produced supernormal formant transitions and extreme vowels at normal speech rates

The Supernormal Plover Egg with Black Spots on a White Background (at left) Preferred by the Plover to the Normal Egg with Dark Brown Spots on a Light Brown Background (at right)



(After Koehler and Zagarus, 1937, reproduced in Tinbergen, 1951.)

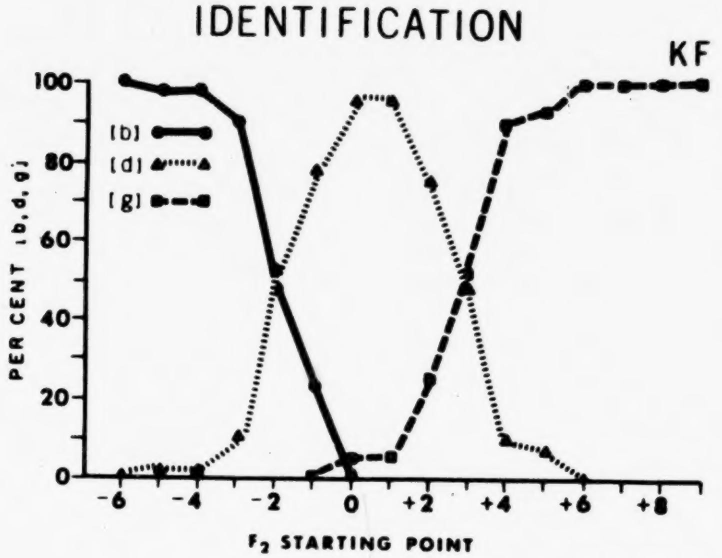
Fig. 5

would probably be unable to perform these primary functions properly. What is more interesting, as Manning goes on to point out, is that the tendency to respond to the sign stimulus has not evolved so as to be perfectly adjusted to the naturally occurring form of the stimulus. Like heterogeneous summation, this must reflect a characteristic of the process by which sign stimuli are perceived, and speech perception must share this characteristic. When we listen to natural speech, presumably we respond best to that combination of cues which approaches the supernormal ideal most closely. Thorpe (1961:98), similarly, has observed that the best natural sign stimulus display is the one which "can come nearest to the supernormal for the largest number of constituent sign stimuli."

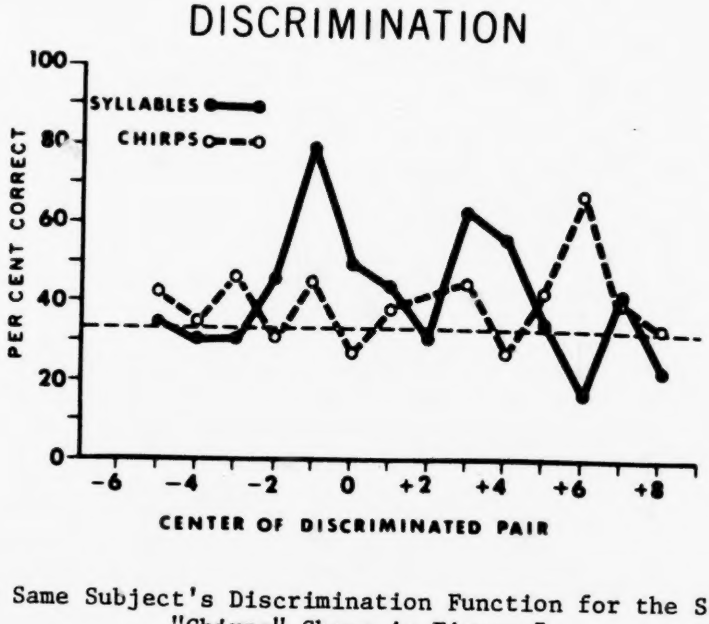
Finally, since the validity of the concept of a specialized neural mechanism to account for the selective perception of and response to sign stimuli is in dispute, the possibility that some such mechanism operates in speech perception is of special interest. The properties which speech perception have in common with sign stimuli point in this direction, for they are not characteristic of human auditory perception in general; so does the possibility of genetic transmission of knowledge of the cues. There is also some other evidence. If we ask a subject to discriminate pairs of stimuli which are adjacent along the acoustic series of stop-vowel syllables with varying F2 transition (Figure 2), he will do very well near the boundaries implied by the cross-over points in his labeling functions and very poorly elsewhere. The upper part of Figure 6 shows the labeling functions of a typical subject; the lower part (solid line) shows his discrimination function for the syllables. He is discriminating categorically (Liberman, 1957). Discrimination of this kind is quite unusual in psychophysical tasks. If we now give the subject a similar discrimination task in which the stimuli are "chirps," i.e., F2 transitions in isolation, without F1 or the steady-state portion of F2 (Figure 7), his discrimination function, represented by the dashed line in the lower part of Figure 6, is quite different. He discriminates better than random for most of the series, but the peaks of the syllable discrimination function are absent. Without a context containing other speech cues, the F2 transition is heard quite differently: there is no indication of categorical perception, and the function is more typically psychophysical (Mattingly et al., 1971).

Additional evidence for a special mechanism comes from experiments in dichotic presentation of speech sounds. If different stop-vowel syllables are simultaneously presented to a subject's two ears, he will be able to report correctly the stimuli presented to the right ear more often than the stimuli presented to the left ear. The effect is attributed to the processing of speech in the left cerebral hemisphere (Kimura, 1961; Studdert-Kennedy and Shankweiler, 1970). No such right-ear advantage is found with nonspeech signals such as musical tones (Kimura, 1964). Experiments by Conrad (1964), Wickelgren (1966), and others suggest that the speech perception mechanism is somehow involved with, and perhaps includes, "short-term memory."

To recapitulate, speech cues have a number of perceptual properties in common with sign stimuli. Their perception does not require a naturalistic context, they obey the law of heterogeneous summation, they are more effective as they approach a supernormal ideal, and there is reason to suppose that a special neural mechanism is involved. Some of these formal properties



Labeling and Discrimination Functions for One Subject for the Series Synthetic Speech Syllables Shown in Figure 2.



The Same Subject's Discrimination Function for the Series of "Chirps" Shown in Figure 7.

Fig. 6

The Pattern for a Series of "Chirps" (Isolated F2 Transitions), Corresponding to the Series of Stop-Vowel Syllables in Figure 2

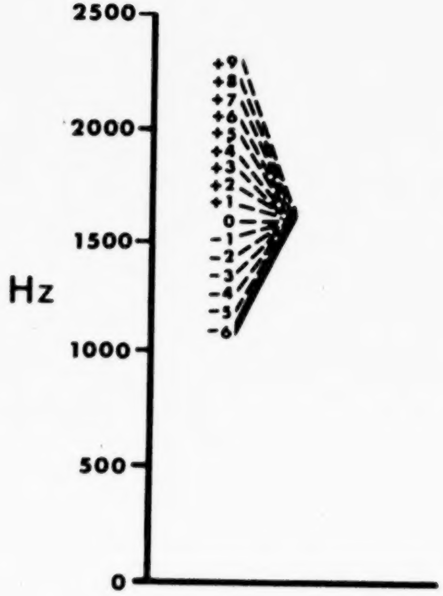


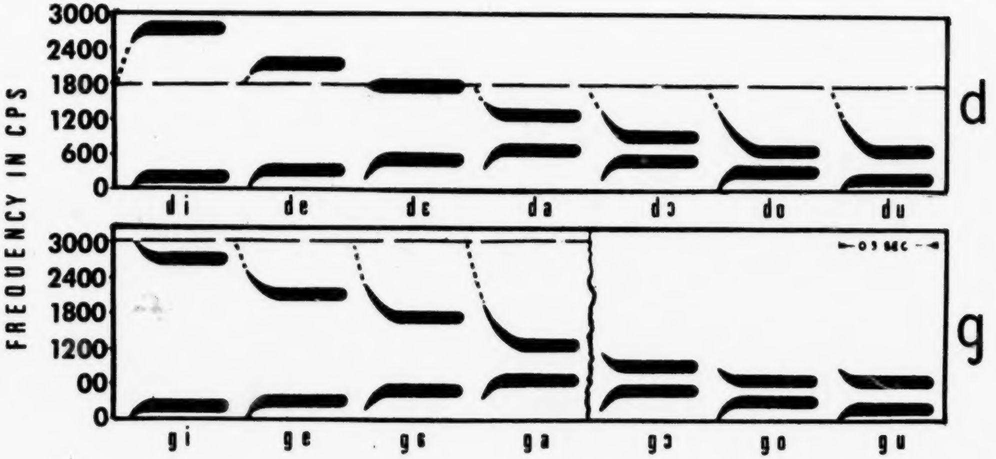
Fig. 7

appear in other situations--heterogeneous summation is a property of human binocular vision for instance--but it is their co-occurrence in both speech and sign stimuli that I find compelling. These properties are shared by the sign stimulus systems of many species, presumably for functional rather than for phylogenetic reasons. Thus, we are led to ask whether speech is in some way functionally similar to a sign stimulus system. But before considering this point, we ought to mention certain rather obvious differences between sign stimuli and the speech cues.

First the speech cues are transmitted at a rate much higher than the sign stimuli of any animal system. The displays in which sign stimuli occur, if not virtually static, are either relatively slow-moving or highly repetitive. But the acoustic events of speech which serve as cues occur extremely rapidly. The speech-perceiving mechanism not only keeps up with these events but is capable, as experiments with speeded speech have demonstrated, of speeds more than three times greater than normal speaking rates (Orr et al., 1965). A further gain in transmission speed is obtained by "parallel processing": the speaker produces and the listener extracts cues for different phonetic distinctions more or less simultaneously from the same acoustic activity (Liberman et al., 1967). Thus in a consonant-vowel syllable, the slope of the transition will carry information about the place of articulation of a consonant, its manner class (stop, fricative, semivowel) and about the quality of the vowel, while the excitation of these same transitions will cue the voicing distinction. The information rate of speech can be as high as 150 bits/second, and the question of the adaptive value of such a high rate arises.

Another difference between speech cues and sign stimuli is implicit in our use thus far of such terms as "place of articulation." Although the speech cues are acoustic events, the phonetic distinctions perceived by the listener are not acoustic but articulatory. Thus, the cues for, say, the alveolar sounds [t,d]--a high-frequency burst, an F2 transition which has a locus at about 1800 Hz, and an F3 transition with a locus at 3200 Hz--seem like a highly arbitrary selection if they are regarded as purely acoustic events. Moreover, the events do not occur synchronously; and, as we have just noted, they are interspersed with cues for other phonetic distinctions. But if these same events are interpreted as acoustic correlates of the simple articulatory gesture which produces [t,d], both the selection of events themselves and their relative timing appears quite straightforward. Another indication of the articulatory reference of the cues is that a series of stimuli may be perceived as belonging to the same phonetic category, even though they are not neighbors on an acoustic continuum, but they must not fail to be close together on some articulatory continuum. Thus the series of stimuli heard as [d] before vowels ordered from high front to low back form both an articulatory and an acoustic continuum, defined (though in somewhat oversimplified fashion) by the [t,d] locus (see the upper portion of Figure 8). But in the case of [k,g] the acoustic continuum is incomplete because the concept of the locus fails to apply consistently; the locus for [k,g] with low back vowels appears to be much lower and less clearly specifiable than for high front vowels (lower portion of Figure 8). Yet the perception is constant because the articulation is similar (Liberman, 1957). Conversely, the series of stimuli in Figure 2, which do form an acoustic continuum, divides into [b,d,g,] because the articulatory reference changes abruptly at

F1 and F2 Patterns for [d] and [g] Followed by a Series of Vowels
from High Front to Low Back



Note the discontinuity in F2 for [g].

Fig. 8

two points on the continuum. Because of such phenomena, it seems reasonable to regard speech as an acoustic encoding of articulatory gestures, or rather of the motor commands underlying those gestures (Lisker et al., 1962; Liberman et al., 1963; Studdert-Kennedy et al., 1970). We may call the sequence of motor commands which determines the speaker's output the "phonetic representation." The listener, because of his intuitive knowledge of the speech code, can recover this representation.

The most notable difference between speech cues and sign stimuli is that while sign stimuli typically produce a stereotyped behavioral response, speech cues do not. The reason the response to speech is not stereotyped is of course that unlike sign stimulus displays, a phonetic representation has no fixed significance apart from the linguistic system in which it functions; in itself it is a meaningless pattern, related only quite indirectly to the semantic values of the speakers and hearers. Speech does not stand by itself; it functions as part of language. The meaning of an utterance and the nature of the ultimate behavioral response depend not just on the characters of the stimulus, the environmental context, and the internal state of the perceiver, but also upon something not found in conjunction with any set of sign stimuli--a grammar. By virtue of a system of grammatical rules, shared by speaker and hearer, the speaker can evoke not just a few stereotyped responses but a wide variety, many of which are delayed or covert, and in principle, an infinite range of semantic values can be expressed. The problem is to explain why and how such a powerful system should have evolved.

It is with this problem that most attempts to find precedents for human language in animal behavior have begun. The cries of animals grossly resembling man, as well as animal communications systems which transmit a substantial amount of information even though the physical nature of the signals may be very different from human speech, have been scrutinized by many investigators for linguistic properties. These efforts have consistently failed. The properties treated as linguistic by some investigators have been so abstract--for example, the Hockett-Altmann "design features" (Altmann, 1967; Hockett and Altmann, 1968)--that those characteristics which distinguish language from purposive behavior in general are lost to view (Chomsky, 1968:60) and really fundamental features are placed on a level with trivial ones. Thus Hockett's Design Feature 3, "Rapid Fading," a property shared by all acoustic phenomena, is apparently just as important as DF 13, "Duality of Patterning," which, as we shall see, is truly significant. It is perhaps noteworthy that, according to Hockett and Altmann, the stickleback's communication system, which is of great interest from the viewpoint adopted here, lacks most of the linguistic Design Features.

Other investigators have tried indiscriminately to force the phenomena of animal behavior into standard linguistic categories. In Lenneberg's (1967: 228) words, they have attempted

> to count the number of words in the language of gibbons, to look for phomemes in the vocalizations of monkeys or songs of birds, or to collect the morphemes in the communication systems of bees and ants. In many other instances no such explicit endeavors are stated, but the underlying faith appears to be the same since much time and effort is spent in teaching parrots, dolphins or chimpanzee infants to speak English.

Such efforts, I think, are doomed to failure, and those who have insisted most strongly on the "biological basis of language"--Chomsky and Lenneberg--share this view. Chomsky (1968:62) suggests that human language "is an example of true 'emergence'--the appearance of a qualitatively different phenomenon at a specific stage of complexity of organization." Lenneberg (1967) believes that language has for the most part evolved covertly. In his view, we cannot expect that the steps in the evolution of a characteristic A from some quite different characteristic B will necessarily be manifest. The nature of the process of genetic modification is such that the intervening steps must in many cases remain obscure. This, he suggests, is the case with human language. While Lenneberg's general position on the nature of evolution may well be essentially correct, to take refuge in this position in the case of a particular evolutionary problem, such as the origin of human language, is essentially to abandon the problem.[1]

Despite the lack of precedents for grammar, I think that Chomsky and Lenneberg are perhaps unduly pessimistic and that the parallels between the speech cues and the sign stimuli suggest some interesting speculations about the origins of language.

One of the traditional explanations of language is that it developed from cries of anger, pain, and pleasure (see, e.g., Rousseau, 1755). The difficulty with this explanation is that it does not attempt to account for the transition from cries to names, or for the emergence of grammar. But let us put these problems to one side for the moment and postulate, just as the traditional explanation does, a stage in man's evolution when speech existed independently of language. Such speech, we suppose, had no syntax or semantics. But it was more than just expressive because it had phonetic structure. Its utterances were phonetic representations encoded by acoustic cues. If we ask what function such prelinguistic but structured speech could have had, the parallels we have discussed between speech cues and sign stimuli suggest a possible answer. Since speech is intraspecific, we suggest that it may have been, at this stage of evolution, a social releaser. If this speculation is correct, prelinguistic speech may have served early man as a vehicle for threat behavior, as a reproductive isolating mechanism, and as a means for mutual recognition of human parents and offspring. By means of phonetic representations underlying his utterances, man elicited appropriate behavioral responses from his fellows in each of these crucial situations. It is probably pointless to speculate as to what particular phonetic representations evoked what responses, but it perhaps reflects the primitive function which we

---

[1] Even if precedents for grammar existed in animal communication, it would be very difficult to learn about them. Most of what we know of the grammatical aspects of human language we know not from observations of human behavior but by virtue of our special status as members of the human species. The work of the linguist depends on the availability to him of the intuitions of speakers of a language that certain utterances are, or are not, grammatical. A member of another species, however intelligent, would find it difficult to deduce the most elementary grammatical concepts by observing and manipulating behavior: he would have, somehow, to consult the grammatical intuitions of a human speaker. We are similarly at a loss when speculating about the possible grammars of animal communication systems.

have attributed to speech that while the segmental aspects of speech have been adapted for linguistic purposes, the prosodic features remain as a primary means of physically harmless fighting, of courting, and of demonstrating and responding to parental affection.

If speech was once a social releaser system, we should expect it to show adaptation in the direction of "communications security." While being as conspicuous as possible on appropriate occasions to conspecific individuals, social releasers should be otherwise as inconspicuous as possible, in particular to prey and to predators. In the case of visual releasers, various camouflaging arrangements are found: outside the courtship period, the stickleback changes the color of his belly to a less noticeable shade and birds hide their brilliant plumage (Tinbergen, 1951). In the case of acoustic releasers, the animal can become silent when this is expedient; the simplicity of this solution is the great advantage of acoustic systems. As for speech, two of the differences we have noted between sign stimuli and speech cues are probably to be interpreted as further adaptations in the direction of security. The rapid rate at which the speech cues can be transmitted means that when necessary, transmissions can be extremely brief, making it so much the more difficult for an enemy to locate the source of the signal. And the fact that the articulatory information conveyed by speech can be perceived only by man means that, from the standpoint of other animals, as Hockett and Altmann (1968) point out, human speech is quite literally a code, concealing not only the phonetic representation but also the fact that there is such a representation and that the speaker is human. Presumably the animals man preyed upon would not have been able to distinguish his speech from the chatter of herbivorous nonhuman primates.

Moreover, if we regard speech as a social releaser system, a natural explanation is available for an old problem. The fact that no other animal except man can speak, not even the primates to whom he is most closely related, has long been a cause for wonder and speculation. But, of course, a social releaser is required, almost by definition, to be species-specific: it must be so if it is to perform its authentication function effectively. It is thus no more surprising that speech should be unique to man than that zigzag dances should be unique to sticklebacks.

Let us now consider how the concept of prelinguistic speech as consisting of a system of phonetic social releasers bears on the problem of the origin of language. Most speculations on this topic suppose that man's unusual intelligence must have been the principal factor in the development of language. The weaker version of this view (which would have been that of many post-Bloomfieldian linguists) assumes that man's intelligence differs from that of animals in degree: he alone is intelligent enough to divide the world into its semantic categories and to recognize their predicative relationships. The structure of his language, insofar as it is not purely a matter of convention, reflects the structure of human experience. The stronger version of this view (which I think it is fair to attribute to Chomsky and his colleagues) assumes that man's intelligence differs in kind from that of other animals and that the structure of language, properly understood, reflects specific properties of the human intellect. Speech, according to either version, serves simply as the vehicle for the abstract structure of language. The anatomy of the vocal tract imposes certain practical constraints on linguistic behavior but has only a trivial relationship to linguistic structure.

The difficulty with this view is not only that it makes no attempt to account for the choice of speech as the vehicle of language, but also that many animals display some degree of intelligence, and a few display intelligent behavior comparable in some ways to man's. One would expect to find some limited linguistic behavior among animals of limited intelligence, or something approximating human linguistic behavior among animals whose intelligence seems to resemble man's. But, as we have seen, precedents of any kind are lacking, and it is argued that language is an instance of evolutionary "emergence."

I wish to suggest a somewhat less drastic alternative to emergence. This is that language be regarded as the result of the fortunate coexistence in man of two independent mechanisms: an intellect, capable of making a semantic representation of the world of experience, and the phonetic social-releaser system, a reliable and rapid carrier of information. From these mechanisms a method evolved for representing semantic values in communicable form.

Before this could happen, a means had to be found for the speaker-hearer to recode semantic representations into phonetic representations, and phonetic representations into semantic representations. Clearly this recoding is a complex process, if only because the intellect, being capable of representing a wide range of human experience, probably has a very large number of categorical features available for semantic representations in long-term memory, while the phonetically significant configurations of the vocal tract can be described in terms of a very small number of categorical features--fifteen or twenty at most (Chomsky and Halle, 1968). It would thus be impossible to accomplish the recoding simply by mapping semantic features onto phonetic features. It was necessary for another mechanism to evolve: linguistic capacity, the ability to learn the grammar of a language.[2] The grammar is a description of the complex but rule-governed relationships, in part universal, in part language-specific, which obtain between semantic representations and phonetic representations. By virtue of his grammatical competence, a person can speak and understand utterances in the language according to the rules of grammar.[3]

---

[2] In this discussion, I have ignored for simplicity's sake the obvious fact that there are not one but many languages, each with its own grammar. To Rousseau (1755) and von Humboldt (1836), to explain the diversity of human languages was regarded as a problem second in importance only to that of explaining the origin of language. Recently, Nottebohm (1970) has offered the intriguing suggestion, based on an analogy with bird song, that language diversity enables some members of a species to develop traits appropriate to their particular environment without an irreversible commitment to subspeciation.

[3] The account of the organization of grammar given here, necessarily oversimplified, is based on Chomsky (1965, 1966).

One component of the grammar is the lexicon, a list of morphemes with which semantic, syntactic, and phonological information is associated. The stock of morphemes in a language is large but finite, while the number of conceivable semantic representations is infinite. But an infinite number of grammatical strings of morphemes can be generated by the syntactic component of the grammar, and from these, the semantic component can generate a correspondingly infinite number of semantic representations. The phonological component parallels the semantic component: for each string of grammatical morphemes, a phonetic representation can be generated. The speaker's task is thus to find a phonetic representation which corresponds grammatically to a given semantic representation, while the hearer's task is to find a semantic representation corresponding to a given phonetic representation. In both his roles, the speaker-hearer, in order to recode, must determine heuristically the probable input to a grammatical component, given its output and the rules which generate output from input. Very little is known about how he performs these tasks.

For our purposes, however, the important point is that a grammar has an obvious symmetry. There is a core, the syntactical and lexical components, and two other components, the semantic and the phonological, which generate the semantic and phonetic representations, respectively. The nature of the semantic component, and the representation it generates, appear to be appropriate for storage in long-term memory. The nature of the phonological component, and the representation _it_ generates, are appropriate for on-line transmission by the vocal tract. To relate these two representations is the main motivation of the grammar, and its form is determined both by the properties of the intellect and by those of the phonetic social-releaser system. It is thus surely not correct to view speech as if it were merely selected by happenstance as a convenient vehicle for language.

Once the grammar had begun to develop, we should not be surprised to find that it exercised a reciprocal influence on the development both of the phonetic system and of the intellect. In the case of the former, it has been argued very persuasively (Lieberman et al., in press; Lieberman and Crelin, 1971) that the vocal tract of modern man has evolved from something rather like that of a chimpanzee to its present form, with a shorter jaw, a wider and deeper pharynx, and vocal cords for which the tension is more finely controlled, and that these modifications not only have no other discernible adaptive value than to increase the reliability and the richness of structure of human speech but are actually disadvantageous for the vocal tract's primary functions of chewing, breathing, and swallowing. If man's vocal tract has evolved in this way, corresponding modifications must have taken place in the neural mechanisms for production and perception of speech, resulting in the speech code in the form we now know it. The evidence for the development and specialization of the human intellect as a result of its grammatical affinities is, of course, far less concrete, but the very least that can be said is that the capability of symbolizing things and ideas by words permits a degree of conceptual abstraction without which the kind of thinking which human beings regularly do would be impossible.

If the function of a grammar is to serve as an interface between the phonetic and semantic domains, it is hardly surprising that precedents for linguistic behavior have not been found. The speech production and perception system is a highly specific mechanism; so also is the human intellect. Their co-occurrence in man was a remarkable piece of luck; other animals, which on behavioral or physiological grounds appear to be of high intelligence, had no opportunity to develop language because they lacked a suitable pre-existing communications system. Moreover, even if high intelligence and an appropriate communications system had co-occurred in some other species and combined to form a "language," its grammar would be utterly different in form from any human grammar, because the intellectual and communicative mechanisms from which it evolved would be quite different in detail from the corresponding human mechanisms. In the circumstances, the most we can hope for is to understand more about the separate evolution of the intellect and that of the speech code and to interpret human grammars in terms of their dual origin.

To summarize, I have called attention to certain parallels between the speech cues and sign stimuli. These parallels suggest the speculation that prelinguistic speech may have functioned as a social-releaser system, which would explain the fact that speech is species-specific. It is suggested, furthermore, that human language is not simply the product of the human intellect but is rather to be viewed as the joint product of the intellect and of this prelinguistic communications system. Grammar evolved to interrelate these two originally independent systems. Its dual origin explains the lack of precedents for language in animal behavior and its apparent "emergence."

REFERENCES

Abramson, A. and Lisker, L. (1970) Discriminability along the voicing continuum: Cross-language tests. In Proc. 6th International Cong. Phonetic Sciences. (Prague: Academia).

Altmann, S.A. (1967) The structure of primate social communication. In Social Communication among Primates, S.A. Altmann, ed. (Chicago: Univ. of Chicago Press).

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: M.I.T. Press).

Chomsky, N. (1966) Topics in the Theory of Generative Grammar. (The Hague: Mouton).

Chomsky, N. (1968) Language and Mind. (New York: Harcourt Brace).

Chomsky, N. and Halle, M. (1968) The Sound Pattern of English. (New York: Harper and Row).

Conrad, R. (1964) Acoustic confusions in immediate memory. Brit. J. Psychol. 55, 75-83.

Cooper, F.S. (1950) Spectrum analysis. J. acoust. Soc. Amer. 22, 761-762.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M. and Gerstman, L.J. (1952) Some experiments on the perception of synthetic speech sounds. J. acoust. Soc. Amer. 24, 597-606.

Delattre, P.C., Liberman, A.M. and Cooper, F.S. (1955) Acoustic loci and transitional cues for consonants. J. acoust. Soc. Amer. 27, 769-773.

Eibl-Eibesfeldt, I. (1970) Ethology. (New York: Holt, Rinehart and Winston)

Eimas, P.D., Siqueland, E.R., Jusczyk, P. and Vigorito, J. (1970) Speech perception in infants. Science 171, 303-306.

Falls, J.B. (1969) Functions of territorial song in the white-throated sparrow. In Bird Vocalizations in Relation to Current Problems in Biology and Psychology, R.A. Hinde, ed. (Cambridge: Cambridge University Press).

Fant, C.G.M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Hailman, J.P. (1969) How an instinct is learned. Scientific American 221, 6, 98-106.

Halle, M., Hughes, G.W. and Radley, J.-P.A. (1957) Acoustic properties of stop consonants. J. acoust. Soc. Amer. 29, 107-116.

Harris, K.S., Hoffman, H.S., Liberman, A.M., Delattre, P.C. and Cooper, F.S. (1958) Effect of third-formant transitions on the perception of the voiced stop consonants. J. acoust. Soc. Amer. 30, 122-126.

Hinde, R.A. (1970) Animal Behavior. 2nd ed. (New York: McGraw-Hill).

Hockett, C.F. and Altmann, S.A. (1968) A note on design features. In Animal Communication, T.A. Sebeok, ed. (Bloomington: Indiana Univ. Press).

Hoffman, H.S. (1958) Study of some cues in the perception of the voiced stop consonants. J. acoust. Soc. Amer. 30, 1035-1041.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. exp. Psychol. 16, 355-358.

Koehler, O. and Zagarus, A. (1937) Beiträge zum Brutverhalten des Halsband-regenpfeifers (Charadrius h. hiaticula L.). Beitr. Fortpflanzungs biol. Vögel 13, 1-9. Cited by Tinbergen, 1951.

Lenneberg, E. (1967) Biological Foundations of Language. (New York: John Wiley).

Liberman, A.M. (1957) Some results of research on speech perception. J. acoust. Soc. Amer. 29, 117-123.

Liberman, A.M., Cooper, F.S., Harris, K.S. (1963) A motor theory of speech perception. In Proceedings of the Speech Communications Seminar. (Stockholm: Speech Transmission Laboratory, Royal Institute of Technology).

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A.M., Delattre, P.C. and Cooper, F.S. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Lang. and Speech 1, 153-167.

Liberman, A.M., Delattre, P.C., Gerstman, L.J. and Cooper, F.S. (1956) Tempo of frequency change as a cue for distinguishing classes of speech sounds. J. exp. Psychol. 52, 127-137.

Lieberman, P., Crelin, E.S. and Klatt, D.H. (in press) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. American Anthropologist.

Lieberman, P. and Crelin, E.S. (1971) On the speech of Neanderthal man. Linguistic Inquiry 2, 203-222.

Lisker, L. (1957) Closure duration and the intervocalic voiced-voiceless distinction in English. Lang. 33, 42-49.

Lisker, L. and Abramson, A.S. (1970) The voicing dimension: Some experiments in comparative phonetics. In Proc. 6th International Cong. Phonetic Sciences. (Prague: Academia).

Lisker, L., Cooper, F.S. and Liberman, A.M. (1962) The uses of experiment in language description. Word 18, 82-106.

Lorenz, K. (1935) Der Kumpan in der Umwelt des Vogels. J. f. Ornith. 83, 137-213, 289-413. Tr. in Instinctive Behaviour, C.H. Schiller, ed. (London: Methuen, 1957).

Lorenz, K. (1954) Das angeborene Erkennen. Natur und Museum 84, 285-295.

Manning, A. (1967) An Introduction to Animal Behavior. (Reading, Mass.: Addison-Wesley).

Mattingly, I.G. (1968) Experimental methods for speech synthesis by rule. IEEE Trans. Audio *16*, 198-202.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K. and Halwes, T. (1971) Discrimination in speech and nonspeech modes. Cognitive Psychol. *2*, 131-157.

Nottebohm, F. (1970) Ontogeny of birdsong. Science *167*, 950-966.

Orr, D.B., Friedman, H.L. and Williams, J.C.C. (1965) Trainability of listening comprehension of speeded discourse. J. educ. Psychol. 56, 148-156.

Rousseau, J.-J. (c.1755) Essay on the origin of languages. Tr. by J. H. Moran in *On the Origin of Language*, J. H. Moran and A. Gode, eds. (New York: Ungar)

Russell, E.S. (1943) Perceptual and sensory signs in instinctive behavior. Proc. Linnacan Soc. London *154*, 195-216.

Seitz, A. (1940) Die Paarbildung bei einigen Cichliden I. Zs. Tierpsychol. *4*, 40-84. Cited in Tinbergen, 1951.

Shearme, J.N. and Holmes, H.N. (1962) An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1-formant 2 plane. In *Proc. Fourth Int. Cong. Phonetic Sciences*. (The Hague: Mouton).

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S. and Cooper, F.S. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. *77*, 234-249.

Studdert-Kennedy, M. and Shankweiler, D. (1970) Hemispheric specialization for speech perception. J. acoust. Soc. Amer. *48*, 579-594.

Thorpe, W.H. (1961) Introduction to: Experimental studies in animal behavior. In *Current Problems in Animal Behaviour*, W.H. Thorpe and O.L. Zangwill, eds. (Cambridge: Cambridge University Press).

Tinbergen, N. (1951) *The Study of Instinct*. (Oxford: Clarendon Press).

von Humboldt, Wilhelm. (1836) *Linguistic Variability and Intellectual Development*. Tr. by G.C. Buck and F.A. Raven. (Coral Gables, Fla.: Univ. of Miami Press, 1971).

Wickelgren, W.A. (1966) Distinctive features and errors in short-term memory for English consonants. J. acoust. Soc. Amer. *39*, 388-398.

On the Evolution of Human Language[*]

Philip Lieberman[+]
Haskins Laboratories, New Haven

ABSTRACT

Recent theoretical and experimental advances have demonstrated that the sounds of human speech make human language an effective medium of communication through a process of speech "encoding." The presence of sounds like the language universal vowels /a/, /u/, and /i/ makes this process possible. In the past five years we have shown that the anatomic basis of human speech is species-specific. We have recently been able to reconstruct the supralaryngeal vocal tracts of extinct hominid species. These reconstructions make use of the methods of comparative anatomy and skeletal similarities that exist between extinct fossils and living primates like newborn _homo sapiens_ and the nonhuman primates. Computer-implemented supralaryngeal vocal tract modelling indicates that these extinct species lacked the anatomic ability that is necessary to produce the range of sounds that is necessary for human speech. Human linguistic ability depends, in part, on the gradual evolution of modern man's supralaryngeal vocal tract. Species like "classic" Neanderthal man undoubtedly had language, but their linguistic ability was markedly inferior to modern man's.

Human language is one of the defining characteristics that differentiate modern man from all other animals. The traditional view concerning the uniqueness of human linguistic ability is that it is based on man's mental processes (Lenneburg, 1967). In other words the "uniqueness" of human language is supposed to be entirely due to the properties of the human brain. The particular sounds that are employed in human language are therefore often viewed as an arbitrary, fortuitously determined set of cipher-like elements. Any other set of sounds or gestures supposedly would be just as useful at the communicative, i.e., the phonetic, level of human language.

The results of recent research have, however, challenged this view. The "motor theory" of speech perception that has been developed over the past fifteen years, in essence, states that speech signals are perceived in terms of the constraints that are imposed by the human vocal apparatus (Liberman et al., 1967). Other recent research, which I will attempt to summarize in this paper, indicates that the anatomic basis of human speech production is itself species-specific. This research is the product of a collaborative effort involving many skills. Edmund S. Crelin of the Yale University School of

---

[*] Paper presented at the Seventh International Congress of Phonetic Sciences, Montreal, 1971. To be published in the Proceedings.

[+] Also University of Connecticut, Storrs.

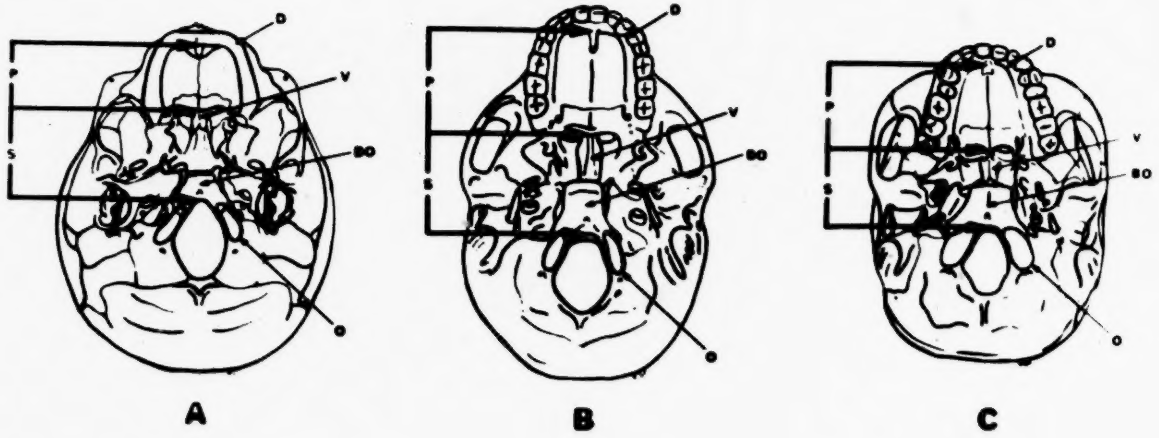
Medicine, Dennis H. Klatt of M.I.T., Peter Wolff of Harvard University, and my colleagues at the University of Connecticut and Haskins Laboratories have all been involved at one time or another. Our research indicates that the anatomic basis of human speech production is the result of a long evolutionary process in which the Darwinian process of natural selection acted to retain mutations that would enhance rapid communication through the medium of speech. The neural processes that are involved in the perception of speech and the unique species-specific aspects of the human supralaryngeal vocal tract furthermore appear to be interrelated in a positive way.

Vocal Tract Reconstruction

The most direct approach to this topic is to start with our most recent experimental technique, the reconstruction and functional modelling of the speech-producing anatomy of extinct fossil hominids. We have been able to reconstruct the evolution of the human supralaryngeal vocal tract by making use of the methods of comparative anatomy and skeletal similarities that exist between extinct fossil hominids and living primates (Lieberman and Crelin, 1971). In Figure 1 inferior views of the base of the skull are shown for newborn modern man, a reconstruction of the fossil La Chappelle-aux-Saints Neanderthal man, and an adult modern man. The detailed morphology of the base of the skull and mandible, which is similar in newborn modern man and Neanderthal man, forms the basis for the Neanderthal reconstruction. Some of the skull features that are similar in newborn modern man and Neanderthal man, but different from adult modern man, are as follows: (1) the skulls have a generally flattened out base; (2) they lack a chin; (3) the body of the mandible is 60 to 100 percent longer than the ramus; (4) the posterior border of the mandibular ramus is markedly slanted away from the vertical plane; (5) there is a more horizontal inclination of the mandibular foramen leading to the mandibular canal; (6) the pterygoid process of the sphenoid bone is relatively short and its lateral lamina is more inclined away from the vertical plane; (7) the styloid process is more inclined away from the vertical plane; (8) the dental arch of the maxilla is U-shaped instead of V-shaped; (9) the basilar part of the occipital bone between the foramen magnum and the sphenoid bone is only slightly inclined away from the horizontal toward the vertical plane; (10) the roof of the nasopharynx is a relatively shallow elongated arch; (11) the vomer bone is relatively short in its vertical height and its posterior border is inclined away from the vertical plane; (12) the vomer bone is relatively far removed from the junction of the sphenoid bone and the basilar side part of the occipital bone; (13) the occipital condyles are relatively small and elongated. These similarities are in accord with other skeletal features typical of Neanderthal fossils (Vlček, 1970), which may be seen in the course of the ontogenetic development of modern man. This, parenthetically, does not mean that Neanderthal man was a direct ancestral form of modern man since Neanderthal fossils exhibit specializations like brow ridges that never occur in the ontogenetic development of modern man. Modern man, furthermore, deviates quite drastically from Neanderthal man in the course of normal maturation from the newborn state.

In Figure 2 lateral views of the skull, vertebral column, and larynx of newborn and adult modern man and Neanderthal man are presented. The significance of the aforementioned skeletal features with regard to the supralaryngeal vocal tract can be seen in the high position of the larynx in newborn and in Neanderthal.

Inferior Views of Base of Skull



NEWBORN NEANDERTHAL ADULT MAN

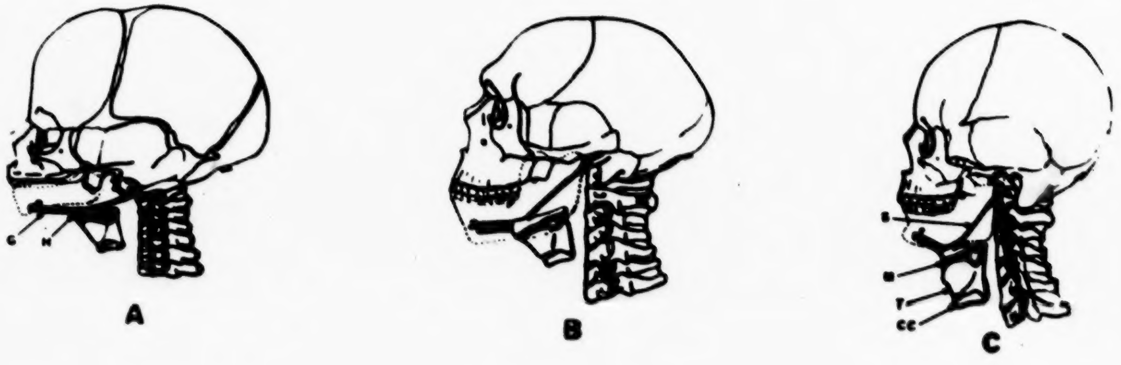
D - Dental Arch
P - Palate
S - Distance Between Palate and Foramen Magnum

V - Vomer Bone
BO - Basilar Part of Occipital
O - Occipital Condyle

(After Lieberman and Crelin, 1971.)

Fig. I

Skull, Vertebral Column, and Larynx



NEWBORN

RECONSTRUCTION OF NEANDERTHAL

ADULT MAN

G - Geniohyoid Muscle
H - Hyoid Bone
S - Stylohyoid Ligament

M - Thyrohyoid Membrane
T - Thyroid Cartilage
CC - Cricoid Cartilage

Fig. 2

Note that the inclination of the styloid process away from the vertical plane in newborn and Neanderthal results in a corresponding inclination in the stylohyoid ligament. The intersection of the stylohyoid ligament and geniohyoid muscle with the hyoid bone of the larynx occurs at a higher position in newborn and Neanderthal. The high position of the larynx in the Neanderthal reconstruction follows, in part, from this intersection. (After Lieberman and Crelin, 1971.)

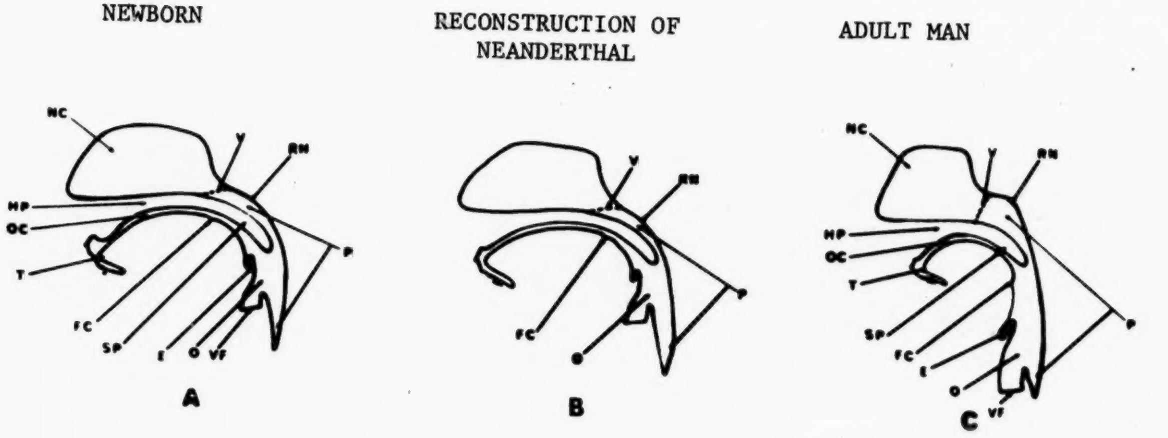
In Figure 3 the supralaryngeal air passages of newborn and adult man and the Neanderthal reconstruction are diagrammed so that they appear equal in size. Although the nasal and oral cavities of Neanderthal are actually larger than those of adult modern man, they are quite similar in shape to those of the newborn. The long "flattened out" base of the skull in newborn and Neanderthal is a concomitant skeletal correlate of a supralaryngeal vocal tract in which the entrance to the pharynx lies behind the entrance to the larynx. In the ontogenetic development of adult modern man the opening of the larynx into the pharynx shifts to a low position. In this shift the epiglottis becomes widely separated from the soft palate. The posterior part of the tongue, between the foramen cecum and the epiglottis, shifts from a horizontal resting position within the oral cavity to a vertical resting position, to form the anterior wall of the oral part of the pharynx (Figure 3C). In this shift the epiglottis becomes widely separated from the soft palate.

The uniqueness of the adult human supralaryngeal vocal tract rests in the fact that the pharynx and oral cavities are almost equal in length and are at right angles. No other animal has this "bent" supralaryngeal vocal tract in which the cross-sectional areas of the oral and pharyngeal cavities can be independently modified. The human vocal tract can, in effect, function as a "two tube" acoustic filter. In Figure 4 we have diagrammed the "bent" human supralaryngeal vocal tract in the production of the "extreme," "point" vowels /i/, /a/, and /u/. Note that the midpoint area function changes are both extreme and abrupt. Abrupt discontinuities can be formed at the midpoint "bend." In Figure 5 the nonhuman "straight" vocal tract which is typical of all living nonhuman primates (Lieberman, 1968; Lieberman et al., 1969, and Lieberman et al., in press), newborn humans (Lieberman et al., 1968), and Neanderthal man, is diagrammed as it approximates these vowels. All area function adjustments have to take place in the oral cavity in the nonhuman supralaryngeal vocal tract. Although midpoint constrictions obviously can be formed in the nonhuman vocal tract, they cannot be both extreme and abrupt. The elastic properties of the tongue prevent it from forming abrupt discontinuities at the midpoint of the oral cavity.

Vocal Tract Modelling

Human speech is essentially the product of a source, the larynx for vowels, and a supralaryngeal vocal tract transfer function. The supralaryngeal vocal tract in effect filters the source (Chiba and Kajiyama, 1958; Fant, 1960). The activity of the larynx determines the fundamental frequency of the vowel, whereas its formant frequencies are the resonant modes of the supralaryngeal vocal tract. The formant frequencies are determined by the area function of the supralaryngeal vocal tract. Man uses his articulators (the tongue, lips, mandible, pharyngeal constrictors, etc.) to modify dynamically in time the formant frequency patterns that the supralaryngeal vocal tract imposes on the speech signal. The phonetic inventory of a language is therefore limited by (1) the number of source function modifications that a speaker is capable of controlling during speech communication and (2) the number of formant frequency patterns available by changing the supralaryngeal area function through the dynamic manipulation of the articulators. We thus can assess the contribution of the supralaryngeal vocal tract to the phonetic abilities of a hominid, independent of the source characteristics. A computer-implemented model of a supralaryngeal vocal tract (Henke, 1966) can be used to determine the possible contribution of the vocal tract to the phonetic repertoire. We can conveniently

Supralaryngeal Air Passages



NC - Nasal Cavity
V - Vomer Bone
RN - Roof of Nasopharynx
P - Pharynx
HP - Hard Palate
SP - Soft Palate
OC - Oral Cavity
T - Tip of Tongue
FC - Foramen Cecum of Tongue
E - Epiglottis
O - Opening of Larynx into Pharynx
VF - Level of Vocal Folds

(After Lieberman and Crelin, 1971.)

Fig. 3

Schematic Diagram of the "Bent" Human Supralaryngeal Vocal Tract

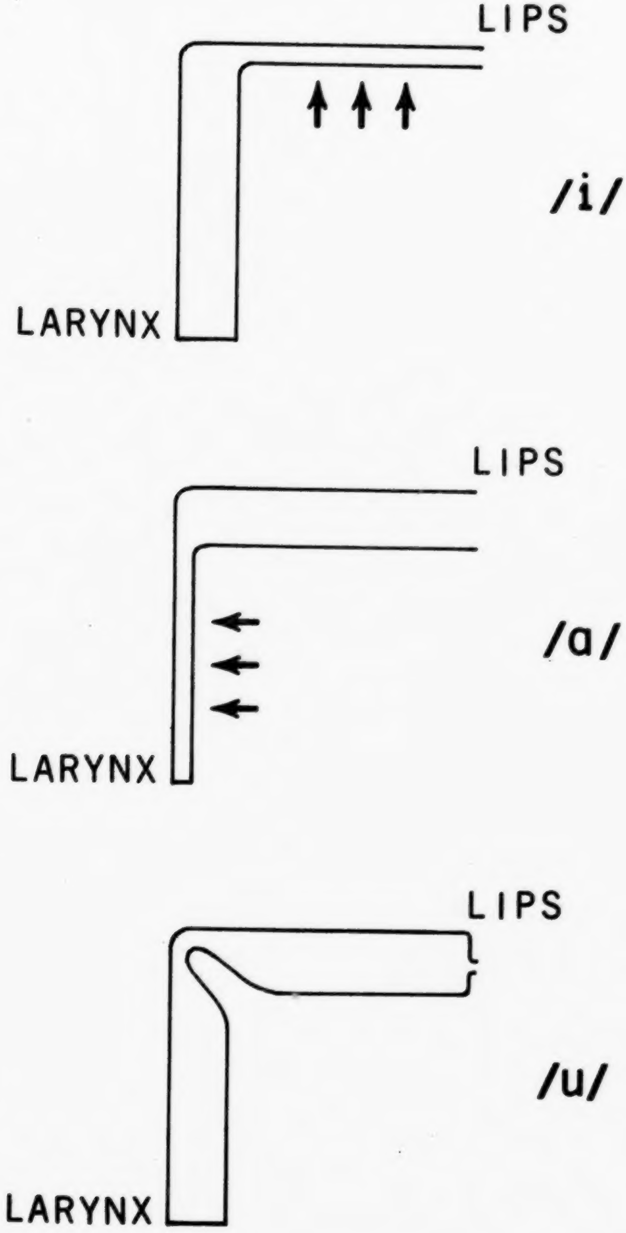


Fig. 4

Note that abrupt and extreme discontinuities in cross-sectional area can occur at the midpoint.

Schematic Diagram of the Straight, "Single Tube" Nonhuman Vocal Tract

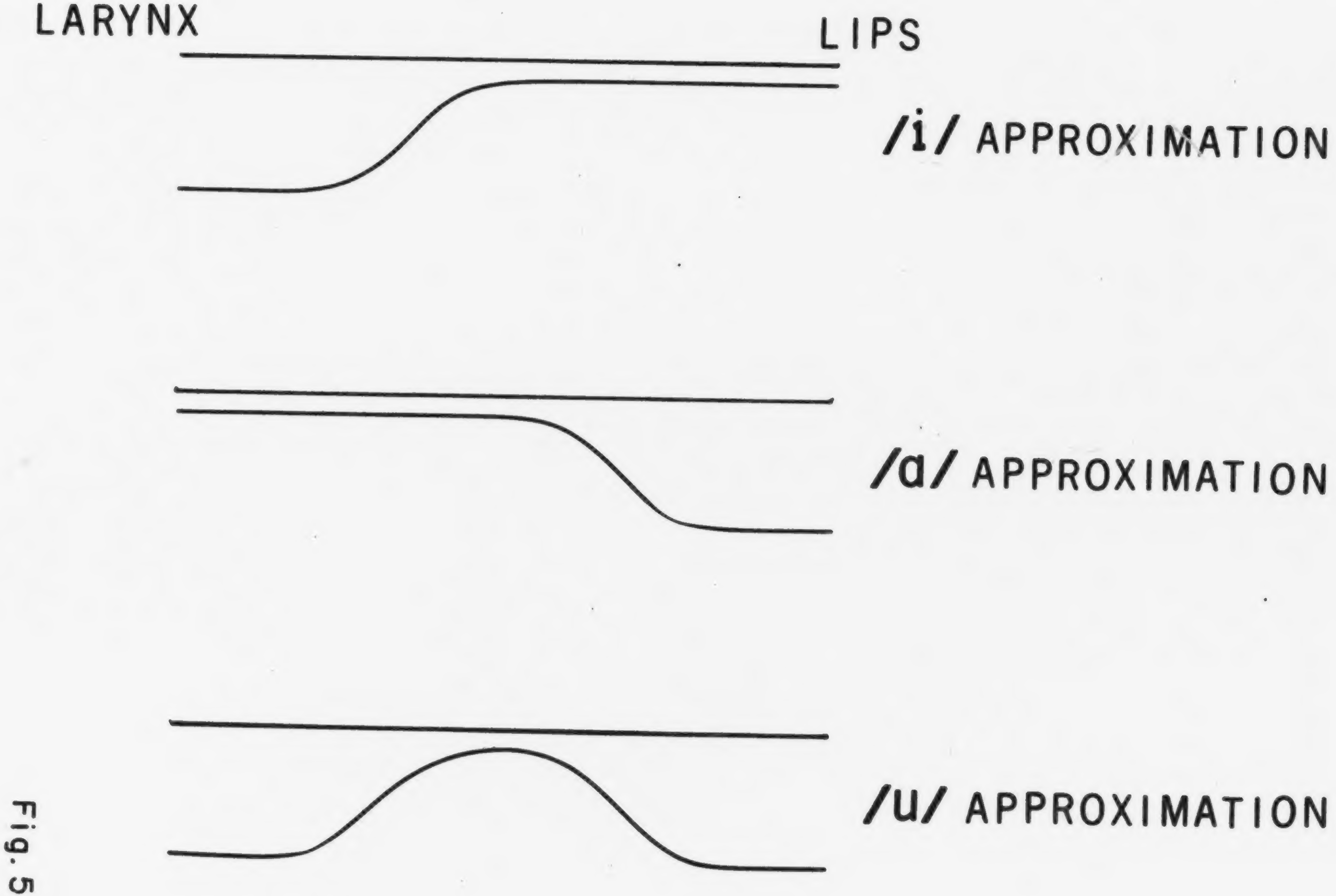


Fig. 5

Note that abrupt midpoint constrictions cannot be formed.

begin to determine whether a nonhuman supralaryngeal vocal tract can produce the range of sounds that occur in human language by exploring its vowel-producing ability. Consonantal vocal tract configurations can also be modelled. It is, however, reasonable to start with vowels since the production of consonants may also involve rapid, coordinated articulatory maneuvers and we can only speculate on the presence of this ability in fossil hominids.

In Figure 6 we have presented area functions of the supralaryngeal vocal tract of Neanderthal man that were modelled on the computer. These area functions were directed towards best approximating the human vowels /i/, /a/, and /u/. Our computer modelling (Lieberman and Crelin, 1971) was guided by the results of X-ray motion pictures of speech production, swallowing, and respiration in adult human (Haskins Laboratories, 1962; Perkell, 1969) and in newborn (Truby et al., 1965). This knowledge plus the known comparative anatomy of the living primates allowed a fairly "conservative" simulation of the vowel-producing ability of classic Neanderthal man. We perhaps allowed a greater vowel-producing range for Neanderthal man since we consistently generated area functions that were more human-like than ape-like whenever we were in doubt. Despite these compensations the Neanderthal vocal tract cannot produce /i/, /a/, or /u/.

In Figure 7 the formant frequency patterns calculated by the computed program for the numbered area functions of Figure 6 are plotted. The labelled loops are derived from the Peterson and Barney (1952) analysis of the vowels of American-English of 76 adult men, adult women, and children. Each loop encloses the data points that accounted for 90 percent of the samples in each vowel category. We have compared the formant frequencies of the simulated Neanderthal vocal tract with this comparatively large sample of human speakers since it shows that the speech deficiencies of the Neanderthal vocal tract are different in kind from the differences that characterize human speakers. Since all human speakers can inherently produce all the vowels of American-English, we have established that the Neanderthal phonetic repertoire is inherently limited. In some instances we generated area functions that would be human-like, even though we felt that we were forcing the articulatory limits of the reconstructed Neanderthal vocal tract (e.g., area functions 3, 9, and 13). However, even with these articulatory gymnastics the Neanderthal vocal tract could not produce the vowel range of American-English.

Functional Phonetic Limitations

There are some special considerations that follow from the absence of the vowels /i/, /a/, and /u/ from the Neanderthal phonetic repertoire. Phonetic analyses have shown that these "point" vowels are the limiting articulations of a vowel triangle that is almost language universal (Troubetzkoy, 1939). The special nature of /i/, /a/, and /u/ can be argued from theoretical grounds as well. Employing simplified and idealized area functions (similar to those sketched in Figure 4) Stevens (1969) has shown that these articulatory configurations (1) are acoustically stable for small changes in articulation and therefore require less precision in articulatory control than similar adjacent articulations and (2) contain a prominent acoustic feature, i.e., two formants that are in close proximity to form a distinct energy concentration.

The vowels /i/, /a/, and /u/ have another unique acoustical property. They are the only vowels in which an acoustic pattern can be related to a

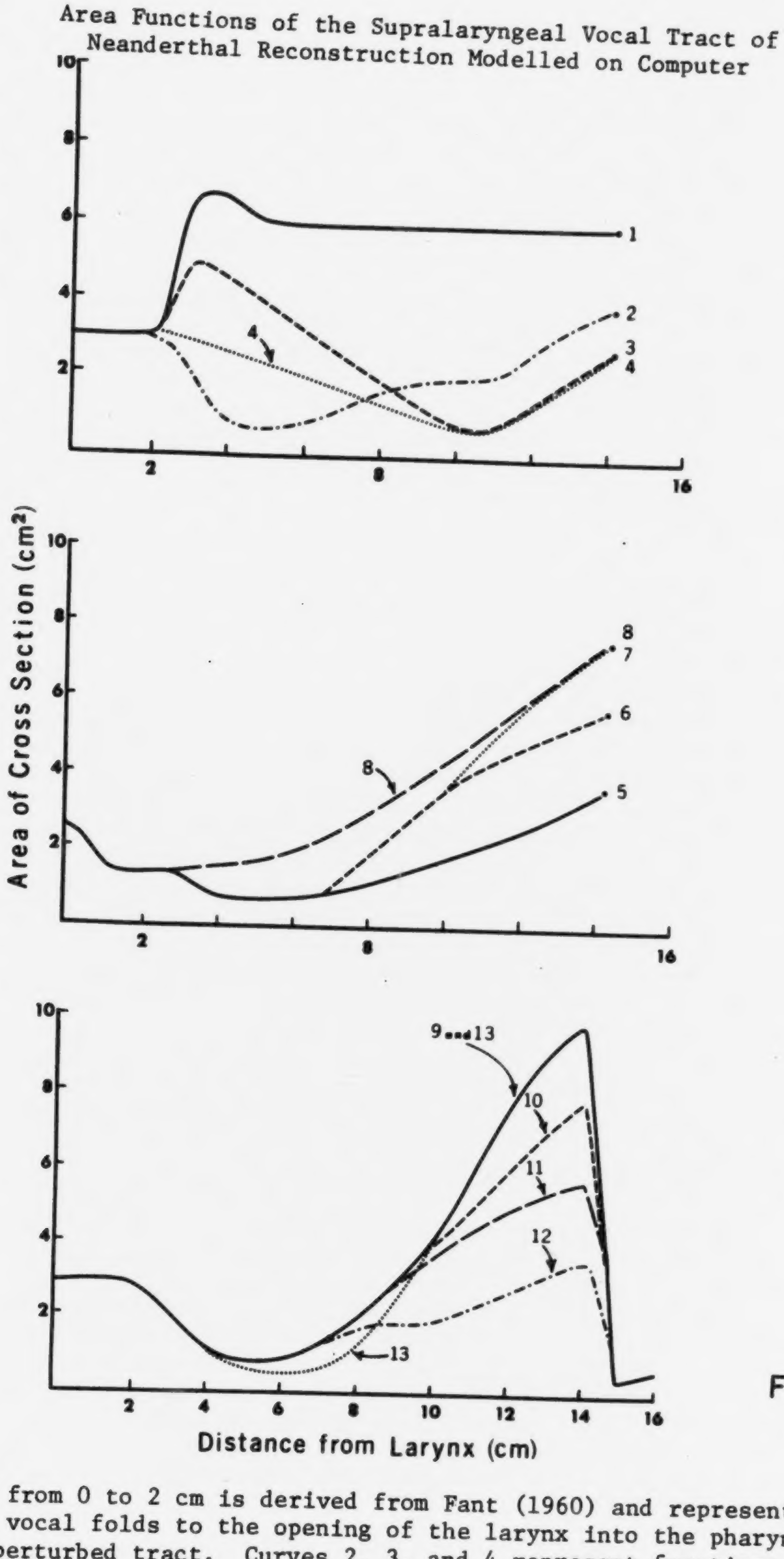
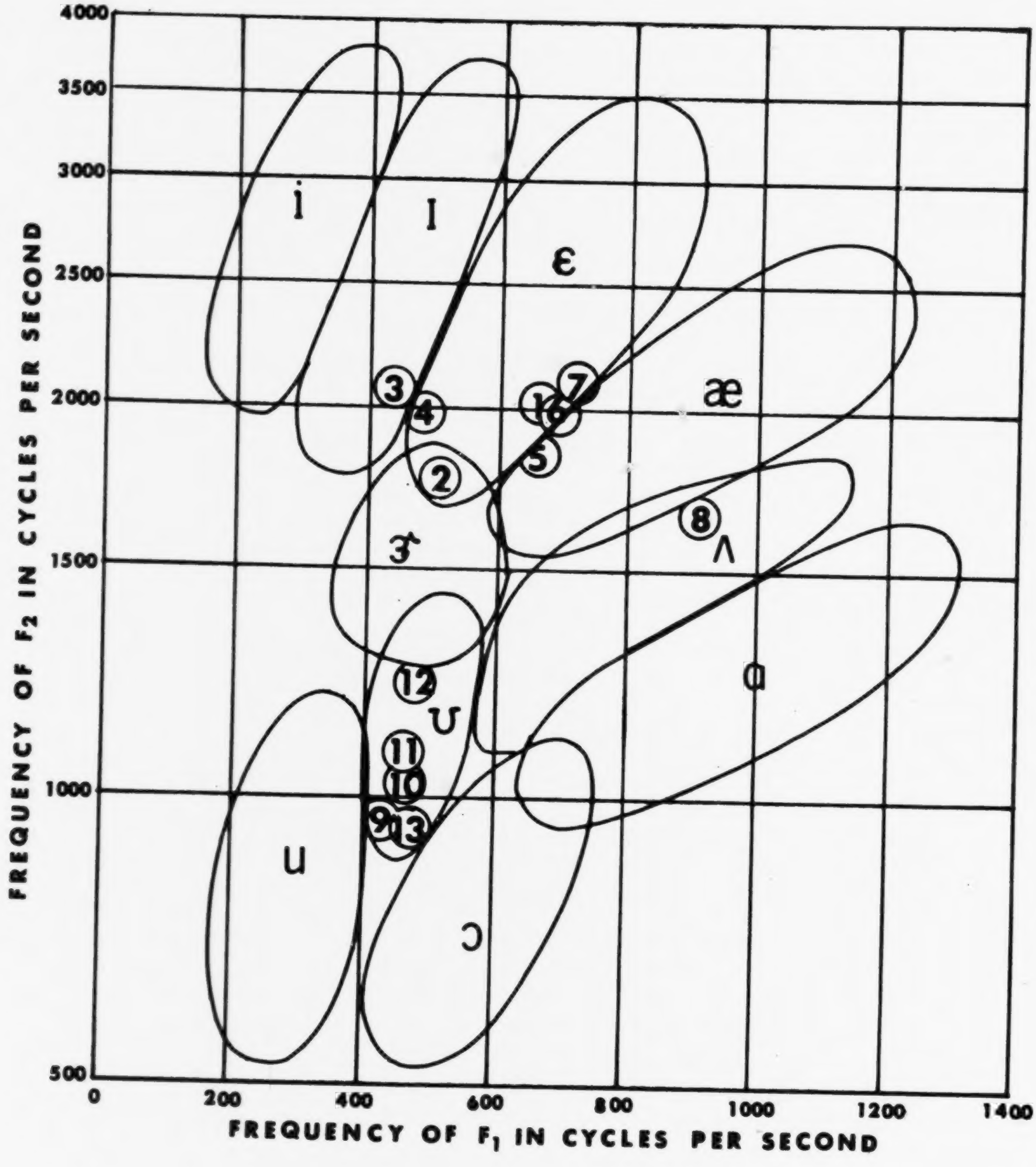


Fig. 6

The area function from 0 to 2 cm is derived from Fant (1960) and represents the distance from the vocal folds to the opening of the larynx into the pharynx. Curve 1 is the unperturbed tract. Curves 2, 3, and 4 represent functions directed towards a "best match" to the human vowel /i/. Curves 5-8 are functions directed towards a "best match" to /a/, while curves 9-13 are directed towards /u/. (After Lieberman and Crelin, 1971.)

Formant Frequencies Calculated by Computer Program for Neanderthal Reconstruction



The numbers refer to area functions in Figure 6. (After Lieberman and Crelin, 1971.)

Fig. 7

unique vocal tract area function. Other "central" vowels can be produced by means of several alternate area functions (Stevens and House, 1955). A human listener, when he hears a syllable that contains a token of /i/, /a/, and /u/, can calculate the size of the supralaryngeal vocal tract that was used to produce the syllable. The listener, in other words, can tell whether a speaker with a large or small vocal tract is speaking. This is not possible for other vowels since a speaker with a small tract can, for example, by increasing the degree of lip rounding, produce a token of /U/ that would be consistent with a larger vocal tract with less lip rounding. These uncertainties do not exist for /i/, /a/, and /u/ since the required discontinuities and constrictions in the supralaryngeal vocal tract area functions produce acoustic patterns that are beyond the range of compensatory maneuvers.

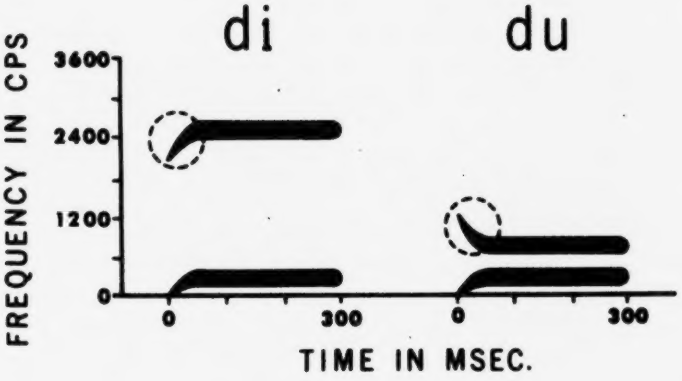
Speech Perception and Speech Anatomy

We noted, at the start of this paper, that the results of perceptual research have demonstrated that human listeners perceive speech in terms of the constraints imposed by the speech-producing apparatus. This mode of perception, which has been termed the "speech" or "motor" theory mode of perception makes the rapid rate of information transfer of human speech possible (Liberman, 1970). Human listeners can perceive as many as 30 phonetic segments per second in normal speech. This information rate far exceeds the temporal resolving power of the human auditory system. It is, for example, impossible even to count simple pulses at rates of 20 pulses per second. The pulses merge into a continuous tone. Human speech achieves its high information rate by means of an "encoding" process that is structured in terms of the anatomic and articulatory constraints of speech production. The motor theory of speech perception, in essence, explicates this process. The presence of vowels like /i/, /a/, and /u/ appears to be one of the anatomic factors that makes this encoding process possible.

In Figure 8 we have reproduced two simplified spectrographic patterns that will, when converted to sound, produce approximations to the syllables /di/ and /du/ (Liberman, 1970). The dark bands on these patterns represent the first- and second-formant frequencies of the supralaryngeal vocal tract as functions of time. Note that the formants rapidly move through a range of frequencies at the left of each pattern. These rapid movements, which occur in about 50 msec, are called transitions. The transition in the second formant, which is encircled, conveys the acoustic information that human listeners interpret as a token of a /d/ in the syllables /di/ and /du/. It is, however, impossible to isolate the acoustic pattern of /d/ in these syllables. If tape recordings of these two syllables are "sliced" with the electronic equivalent of a pair of scissors, it is impossible to find a segment that contains only /d/. There is no way to cut the tape so as to obtain a piece that will produce /d/ without also producing the next vowel or some reduced approximation to it.

Note that the encircled transitions are different for the two syllables. If these encircled transitions are isolated, listeners report that they hear either an upgoing or a falling frequency modulation. In context, with the acoustic correlates of the entire syllable, these transitions cause listeners to hear an "identical" sounding /d/ in both syllables. How does a human listener effect this perceptual response?

Simplified Spectrographic Patterns
Sufficient to Produce the Syllables /di/ and /du/



The circles enclose the second formant frequency transitions. (After Liberman, 1970.)

Fig. 8

We have noted the formant frequency patterns of speech reflect the resonances of the supralaryngeal vocal tract. The formant patterns that define the syllable /di/ in Figure 8 thus reflect the changing resonant pattern of the supralaryngeal vocal tract as the speaker moves his articulators from the occlusion of the tongue tip against the palate that is involved in the production of /d/ to the vocal tract configuration of the /i/. A different acoustic pattern defines the /d/ in the syllable /du/. The resonances of the vocal tract are similar as the speaker forms the initial occlusion of the /d/ in both syllables; however, the resonances of the vocal tract are quite different for the final configurations of the vocal tract for /i/ and /u/. The formant patterns that convey the /d/ in both syllables are thus quite different since they involve transitions from the same starting point to different end points. Human listeners "hear" an identical initial /d/ segment in both of these signals because they "decode" the acoustic pattern in terms of the articulatory gestures and the anatomical apparatus that is involved in the production of speech. The listener in this process, which has been termed the "motor theory of speech perception" (Liberman et al., 1967), operates in terms of the acoustic pattern of the entire syllable. The acoustic cues for the individual "phonetic segments" are fused into a syllabic pattern. The high rate of information transfer of human speech is thus due to the transmission of acoustic information in syllable-sized units. The phonetic elements of each syllable are "encoded" into a single acoustic pattern which is then "decoded" by the listener to yield the phonetic representation.

In order for the process of "motor theory perception" to work the listener must be able to determine the absolute size of the speaker's vocal tract. Similar articulatory gestures will have different acoustic correlates in different-sized vocal tracts. The frequency of the first formant of /a/, for example, varies from 730 to 1030 Hz in the data of Peterson and Barney (1952) for adult men and children. The frequencies of the resonances that occur for various consonants likewise are a function of the size of the speaker's vocal tract. The resonant pattern that is the correlate of the consonant /g/ for a speaker with a large vocal tract may overlap with the resonant pattern of the consonant /d/ for a speaker with a small vocal tract (Rand, 1971). The listener therefore must be able to deduce the size of the speaker's vocal tract before he can assign an acoustic signal to the correct consonantal or vocalic class.

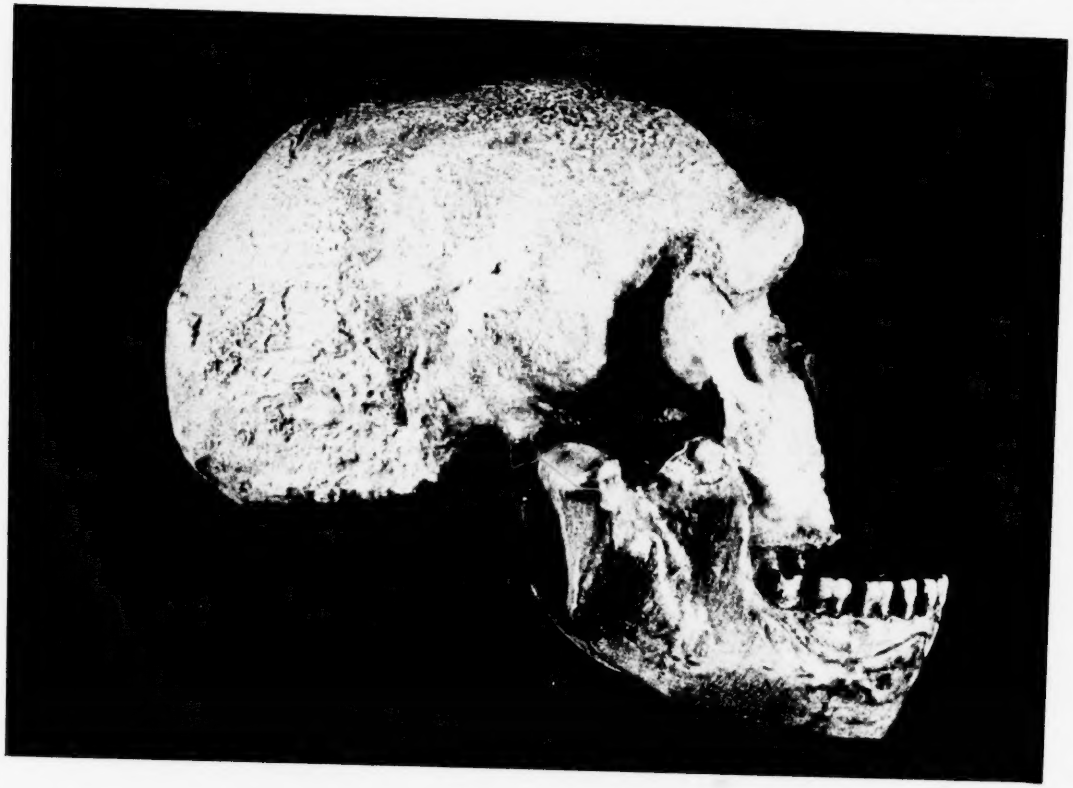
There are a number of ways in which a human listener can infer the size of a speaker's supralaryngeal vocal tract. He can, for example, note the fundamental frequency of phonation. Children, who have smaller vocal tracts, usually have higher fundamental frequencies than adult men or adult women. Adult men, however, have disproportionately lower fundamental frequencies than adult women (Peterson and Barney, 1952), so fundamental frequency is not an infallible cue to vocal tract size. Perceptual experiments (Ladefoged and Broadbent, 1957) have shown that human listeners can make use of the formant frequency range of a short passage of speech to arrive at an estimate of the size of a speaker's vocal tract. Recent experiments, however, show that human listeners do not have to defer their "motor theory" decoding of speech until they hear a two- or three-second interval of speech. Instead, they use the vocalic information encoded in a syllable to decode the syllable (Darwin, in press; Rand, 1971). This may appear to be paradoxical, but it is not. The listener makes use of the formant frequencies and fundamental frequency of the syllable's vowel to assess the size of the vocal tract that produced the syllable. We have noted throughout this paper that the vowels /a/, /i/, and /u/ have a unique acoustical property. The formant frequency pattern for these vowels can always be related to a unique vocal tract size and shape. A listener, when he hears one of these vowels, can thus instantly determine the size of the speaker's vocal tract. The vowels /a/, /i/, and /u/ (and the glides /y/ and /w/) thereby serve as acoustic calibration signals in human speech.

The absence of a human-like pharyngeal region in apes, newborn man, and Neanderthal man is quite reasonable. The only function that the human supralaryngeal vocal tract is better adapted to is speech production, in particular the production of vowels like /a/, /i/, and /u/. The human supralaryngeal vocal tract is otherwise less well adapted for the primary vegetative functions of respiration, chewing, and swallowing (Lieberman et al., 1971; Crelin et al., forthcoming). This suggests that the evolution of the human vocal tract which allows vowels like /a/, /i/, and /u/ to be produced and the universal occurrence of these vowels in human languages reflect a parallel development of the neural and anatomic abilities that are necessary for language. This parallel development would be consistent with the evolution of other human abilities. The ability to use tools depends, for example, both on upright posture and an opposable thumb, and on neural ability.

Neanderthal man lacked the vocal tract that is necessary to produce the human "vocal tract size-calibrating" vowels /a/, /i/, and /u/. This suggests that the speech of Neanderthal man did not make use of syllabic encoding. While communication is obviously possible without syllabic encoding, studies of alternate methods of communication in modern man show, as we noted before, that the rate at which information can be transferred is about one-tenth that of normal human speech.

It is imperative to note that classic Neanderthal man, as typified by fossils whose skull bases are similar to the La Chapelle-aux-Saints, La Ferrasie, La Quina, Pech-de-L'Azé, and Monte Circeo fossil hominids (as well as many others), probably does not represent the mainstream of human evolution. Although Neanderthal man and modern man probably had a common ancestor, Neanderthal represents a divergent species (Boule and Vallois, 1957; Vlček, 1970; Lieberman and Crelin, 1971). In Figure 9 we have photographed a casting of a reconstruction of the fossil Steinheim calvarium with the mandible of the La Chapelle-aux-Saints fossil. The mandible of the Steinheim fossil hominid never was found. Note that the La Chapelle-aux-Saints mandible is too long. In Figure 10 the Steinheim fossil has been fitted with a mandible from a normal adult human, which best "fits" the Steinheim fossil. We are in the process of reconstructing the supralaryngeal vocal tract of the Steinheim fossil (Crelin et al., forthcoming). It is quite likely that this fossil, which is approximately 300,000 years old, had the vocal tract anatomy that is necessary for human speech. The evolution of the anatomical basis for human speech thus would not appear to be the result of abrupt, recent change in the morphology of the skull and soft tissue of the vocal tract. We have noted a number of fossil forms that appear to represent intermediate stages in the evolution of the vocal tract. Recent fossil discoveries indicate that the evolution of the human vocal tract may have started at least 2.6 million years ago. It, therefore, is not surprising to find that the neural aspects of

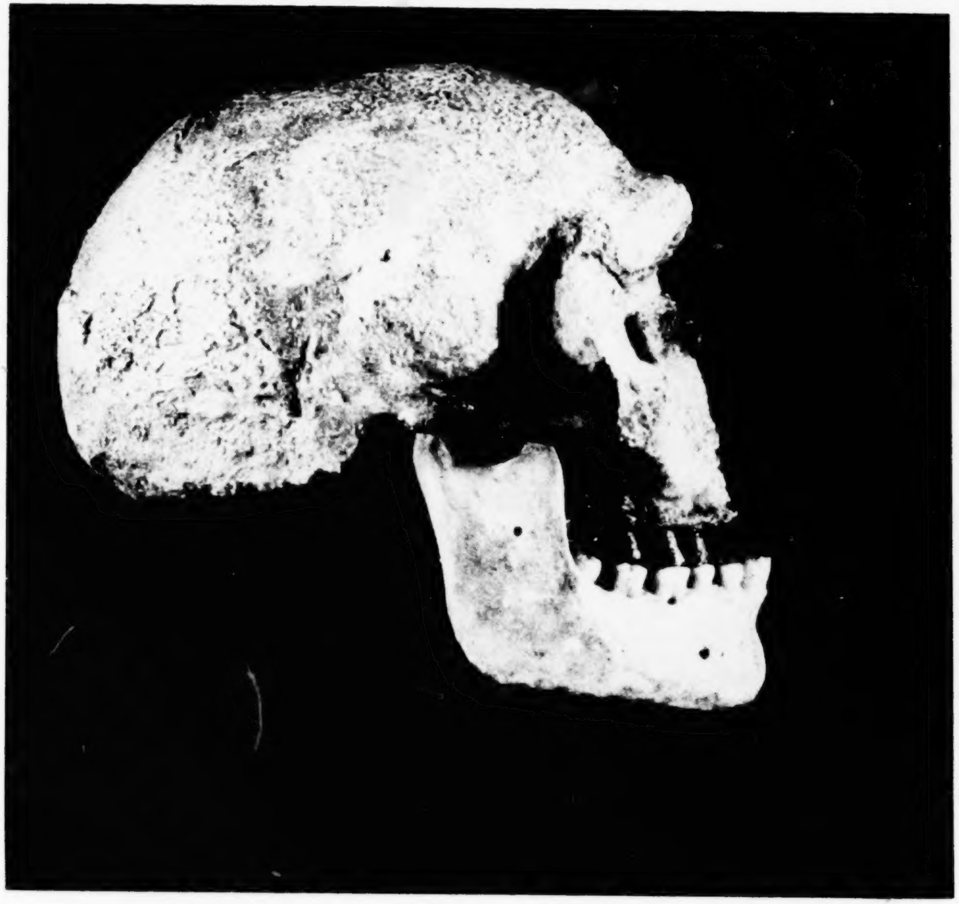
Reconstructed Steinheim Clavarium with Neanderthaloid Mandible



Note that the Neanderthal mandible is too large. (After Crelin et al., forthcoming.)

Fig. 9

Reconstructed Steinheim Clavarium with a Modern Human Mandible



This represents the best "fit." (After Crelin et al., forthcoming.)

Fig. 10

speech perception are matched to the anatomical aspects of speech production. Nor should we be surprised to note that "naturalness" constraints relate the phonetic and phonologic levels of grammar (Jakobson et al., 1952; Postal, 1968; Chomsky and Halle, 1969).

Sir Arthur Keith many years ago speculated on the antiquity of man. We now know that hominid evolution can be traced back at least 3 million years. The evolution of phonetic ability appears to have been an integral part of this evolutionary process. It may have its origins at the very beginnings of hominid evolution.

REFERENCES

Boule, M. and H.V. Vallois (1957) Fossil Men. (New York: Dryden Press).

Chiba, T. and M. Kajiyama (1958) The Vowel: Its Nature and Structure. (Tokyo: Phonetic Society of Japan).

Chomsky, N. and M. Halle (1968) The Sound Pattern of English. (New York: Harper).

Crelin, E.S., P. Lieberman and Dennis Klatt (forthcoming) Speech abilities of the Steinheim, Skhul V, and Rhodesian fossil hominids.

Darwin, C. (in press) Ear differences in the recall of fricatives and vowels. Quarterly J. exp. Psychol. 23.

Fant, G. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Haskins Laboratories (1962) X-Ray Motion Pictures of Speech (New York: Haskins Laboratories).

Henke, W.L. (1966) Dynamic articulatory model of speech production using computer simulation. Unpublished doctoral dissertation, MIT.

Jakobson, R., C.G.M. Fant and M. Halle (1952) Preliminaries to Speech Analysis (Cambridge: M.I.T. Press).

Ladefoged, P. and D.E. Broadbent (1957) Information conveyed by vowel. J. acoust. Soc. Am. 29, 98-104.

Lenneburg, E.H. (1967) Biological Foundations of Language (New York: Wiley).

Liberman, A.M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Liberman, A.M., D.P. Shankweiler, and M. Studdert-Kennedy (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. acoust. Soc. Am. 44, 1574-1584.

Lieberman, P., E.S. Crelin and D.H. Klatt (in press) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man and the chimpanzee. American Anthropologist.

Lieberman, P. and E.S. Crelin (1971) On the speech of Neanderthal man. Linguistic Inquiry 2, 203-222.

Lieberman, P., Dennis H. Klatt and W.A. Wilson (1969) Vocal tract limitations of the vocal repertoires of Rhesus monkey and other non-human primates. Science 164, 1185-1187.

Lieberman, P., K.S. Harris, P. Wolff, and L.H. Russel (1968) Newborn infant cry and non-human primate vocalizations. Haskins Laboratories Status Report on Speech Research 17/18, 23-39.

Perkell, J.S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study (Cambridge, Mass.: MIT Press).

Peterson, G.E. and H.L. Barney (1952) Control methods used in a study of the vowels. J. acoust. Soc. Am. 42, 175-184.

Postal, P.M. (1968) *Aspects of Phonological Theory*. (New York: Harper and Row).

Rand, T.C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research *25/26*, 141-146.

Stevens, K.N. (1969) The quantal nature of speech: Evidence from articulatory-acoustic data. In *Human Communication: A Unified View*, E.E. David, Jr. and P.B. Denes, eds. (New York: McGraw Hill).

Stevens, K.N. and A.S. House (1955) Development of a quantitative description of vowel articulation. J. acoust. Soc. Am. *27*, 484-493.

Troubetzkoy, N.S. (1939) *Principes de phonologie*, Trans., 1949, by J. Cantineau. (Paris: Klincksieck).

Truby, H.M., J.F. Bosman, and J. Lind (1965) *Newborn Infant Cry* (Uppsala: Almqvist and Wiksells).

Vlček, E. (1970) Etude comparative onto-phylogénétique de l'enfant du Pech-de-L'Azé par rapport à d'autres enfants Néandertaliens. In *L'Enfant Du Pech-de-L'Azé*, Ferembach et al. Memoire 33, Archives de L'Institut de Paléontologie Humaine. (Paris: Masson et Cie).

Distinctive Features and Laryngeal Control[*]

Leigh Lisker[+] and Arthur S. Abramson[++]
Haskins Laboratories, New Haven

ABSTRACT

Physiological, acoustic, and perceptual data indicate that the timing of events at the glottis relative to articulation differentiates homorganic stops in many languages. Such categories are variously described in terms of voicing, aspiration, and force of articulation. N. Chomsky and M. Halle have recently proposed a universal set of phonetic features. Four of them--voice, tensity, glottal constriction, and heightened subglottal pressure--allegedly operate to control the onset timing of laryngeal pulsing. Not only is the observational basis for their analysis flimsy, but Chomsky and Halle can advance no substantive argument for rejecting the possibility of temporal control of laryngeal function.

Up until fairly recently the nonhistorical study of language was, at least in this country, pretty much the province of two groups of people: the grammarians and the phoneticians. And it could be said that each group paid little if any serious attention to the problems and findings of the other, even in the area of phonology, where their interests would seem to converge. In the case of the phoneticians, their ignorance of linguistics was not normally elevated to a matter of principle. Some grammarians, however, refused to consider phonetic research an integral part of linguistics. Such work was consigned to physiology and physics at the very time that the primacy of the spoken over the written forms of language was being asserted most emphatically.[1] The dichotomy drawn between _langue_ and _parole_ may have served as an excuse for minimizing the attention given to language in its most directly observable manifestation. Moreover linguists proceeded from the principle that only message-differentiating phonetic features are relevant to language description to the practice of knowing only as much about the processes of speech production and perception as sufficed to provide a set of labels by which to spell different messages distinctively.[2] Insofar as the linguist's concern with the components of sentences and their arrangements is not primarily for the acoustic cues to their recognition, his neglect of phonetics as a serious enterprise may well be justifiable. But for some reason a quite untenable argument has sometimes been advanced--namely that because one cannot hope to achieve a complete and perfectly accurate phonetic description, it follows that no scientific status can be accorded to phonetics. Bloomfield (1933:127-8) aimed this objection at what he called "zealous phonetic experts," and pending the day when phonetic descriptions with a proper degree of trustworthiness would come from a laboratory phonetics of the future, he defined as an adequate phonetic representation the simple encipherment of the phonemes said to make up an utterance. In a very recent statement on phonetics and phonology, Chomsky and Halle's Sound Pattern of English (1968), one of whose merits is its insistence on the need for exposing the nature of the connections among phonetics, phonetic transcription, and phonology, the notion of phonetic transcription as "a device for recording facts observed in actual utterances" (1968:293) is rejected on a basis that seems very like Bloomfield's: transcribers allegedly fail to note everything that a physical recording captures, and they report items for which no physical correlates are found. Of course, add both Bloomfield and Chomsky-Halle, even a perfected phonetic knowledge and a completely faithful transcription would be of doubtful value to the linguist. Bloomfield must also know which physical properties are used by speakers in understanding and repeating utterances, while Chomsky and Halle emphasize the linguist's concern with the "structure of language rather than with the acoustics and physiology of speech" (1968:293). In any case American linguists have seemed happy on the whole to be excused from phonetics class, though they have not refrained from claiming to know a good deal about the articulatory basis for the differences by which utterances are distinguished. Such claims, the judgments of observers with broad experience in listening to varied languages, appear to have merited more respectful attention than had the observations of Bloomfield's "zealous phonetic experts," being apparently immune to the accusation that they might be just as haphazard and just as liable to error. The linguist's phonetics may indeed be more plausible because it is generally less ambitious in the number of distinctions it draws; in point of fact, however, for those distinctions drawn the rather strong claim is made that these are precisely the ones that the native speaker responds to in interpreting the utterances of his language. Both the zealous phonetician's and the linguist's recordings are opinions requiring some sort of control if their scientific status is to be established; the latter, in particular, call for a validation method that involves observation of native listener behavior (Lisker et al., 1962). In either case these recordings, once determined to reflect stable response patterns by the observers

---

[*] Paper to appear in _Language_ (December, 1971). This is a considerably expanded and revised version of the text of an oral paper that appeared under the same title in SR 15/16.

[+] Also University of Pennsylvania, Philadelphia.

[++] Also University of Connecticut, Storrs.

[1] This point has been discussed at length with reference to various linguists by Einar Haugen (1951).

[2] Phoneticians, often enough scolded for doing research not immediately relatable to the linguist's own interests, have generally tried to remedy this situation, but sometimes this seems to take the form of renouncing research in any area that is not directly relevant to linguistics as most narrowly defined. Thus a phonetician with some training in linguistics can write, in connection with a

---

study of mechanical pressures developed in the articulation of certain consonants, that "the nasals are still another matter, as they do not enter into the lenis/fortis opposition, and calculating percentages of overlapping of their values with those of the stops would be meaningless" (Malécot 1966a:176). The fact that phoneticians have failed to exploit research possibilities that closer attention to linguists' discussions would have made them aware of cannot be taken to imply that areas of phonetic research with which linguists have not concerned themselves are therefore without relevance to linguistics. Recent discussion by Mattingly and Liberman (1969) would suggest that linguists have been sometimes too ready to deny linguistic relevance to language and speech studies which threatened to yield findings not readily expressible in the current mode of linguistic description.

to the utterances represented, have still to be matched against physical observations if physical meanings are to be attached to them. Otherwise, at best, the physical features alleged to differentiate utterances are no more than names for classes of impressionistic categories.[3]

Phonetic description and representation, whether to characterize physical regularities in speech behavior or more narrowly to serve as a basis for classifying or spelling utterances, invariably imply the notion of a segment and the specification of segments relative to a finite set of independent or almost-independent dimensions. Current talk of a "universal phonetics" should not obscure the fact that there really is no other. Perhaps the older phonetics is only less prone to claiming that the known set of phonetic dimensions is the set of all possible ones. The point to the recent escalation in the sweep of assertions as to the completeness of our present knowledge is perhaps more than rhetorical; presumably new dimensions or features are not lightly admitted to serious consideration, and the enlargement of the universal set is a properly dramatic event. Of course phonetics, now universal phonetics, is concerned with more than the enumeration and physical specification of features; it has to do also with the nature of their interrelations as determined by universal constraints on speech production and perception. Thus Chomsky and Halle (1968: 294-5) make the strong claim that the features of their universal phonetics are not only components of a labelling system, in which function they have the well-known abstract binary property, but that they also represent, in concrete multivalued fashion, the speech-producing capabilities of the human vocal tract. It is in this latter guise, where speech representation and underlying phonetic assertions are given physical interpretations, that we are concerned with the distinctive features as these are described in the seventh chapter of The Sound Pattern of English.[4]

For some time we have been collecting various kinds of data bearing on the dimension of voicing, or glottal pulsing, as an attribute of initial stop consonants, our aim being to determine in detail just how a single dimension is exploited in a number of languages. Such data, we felt, would be relevant to the general concern of physical phonetics for exploring questions of the following kinds. (1) To what extent is it possible to correlate the phonetic dimensions with which the linguist operates, and for which he claims a distinctive function in particular languages, with measurements of physical properties usually connected with those dimensions? (2) Do languages agree sufficiently in the way in which they divide a dimension into subranges to justify our talking about universal categories? (3) Does a given phonetic dimension interact with others in ways that are not merely language-specific? There were several reasons for focusing attention on glottal pulsing and initial stops. Voicing has a generally agreed-upon acoustic correlate that is readily visible in spectrograms and other displays of the speech signal; voicing differences seem to be widely used in languages to separate stop categories; and voicing is said to co-occur frequently with certain other features, especially with aspiration and differences in what is called "force of articulation." The measure we used was one of the timing of voice onset relative to the release of stop occlusion (Lisker and Abramson, 1964). This measure is most easily applied to stops in utterance-initial position, and we began with those, later extending our observations to other positions as well. Our measurements suggested a number of generalizations: (1) Differences in the relative timing of voice onset show a high correlation with some of the manner distinctions among the stop categories within many languages. (2) By and large, there is rough agreement across languages in the placement of category boundaries along the dimension of voice-onset timing, yielding the three phonetic types: voiced, voiceless unaspirated, and voiceless aspirated stops. (3) The timing of voice onset is somewhat affected by certain contextual factors, among these the place of stop articulation, position in isolated words as against longer stretches of speech, and, for English at least, position relative to degree of syllable prominence (Lisker and Abramson, 1967). These data, derived from samples of a dozen languages, were supplemented by data from experiments in the perception of synthetic speech (Abramson and Lisker, 1965, 1970), by records of intraoral air pressures developed during production of the English stops (Lisker, 1970), and by data from transillumination (Lisker et al., 1969) and fiberoptics photography of the larynx (Sawashima et al., 1970). In addition, there were available mechanical pressure data from Malécot (1966a), and electromyographic data from Harris, Lysaught, and Schvey (1965), Fromkin (1966), and Tatham and Morton (1969). Indeed, Lubker and Parris (1970) made simultaneous intraoral air pressure, mechanical pressure, and electromyographic recordings. All these data have led us to suppose that it is primarily in their control of the timing of laryngeal adjustments relative to supraglottal gestures, rather than differences among those supraglottal gestures, that speakers manifest their choice from a set of homorganic stop categories. It is at present an open question as to whether the speech mechanism is inherently capable of producing stops whose variability in respect to voice-onset timing is essentially continuous over the entire range for which values have been recorded, but limited data derived from mimicry experiments suggest that there is no purely mechanical constraint on such a capability. To say, as we have, that voice-onset timing is the single most effective measure whereby homorganic stop categories in languages generally may be distinguished physically and perceptually does not imply that no other measure need ever be applied in the case of some particular language. Nor do we mean to assert that the speaker's alleged control over voice timing is necessarily exerted in the simplest, most straightforward manner; it might be a matter of varying the time of arrival of neural motor signals to the appropriate laryngeal muscles to close the glottis, but it might as well involve complex changes in the balance of forces exerted by the various muscles acting in and upon the larynx. Moreover, since adjustments elsewhere in the vocal tract are known to affect the operation of the larynx, we cannot rule out the possibility that one or more of these play a significant role in effecting category distinctions. What we do maintain is that for many languages such extralaryngeal adjustments serve primarily to control voice timing, and that any single measure based on one of these features is less useful than the one of voice timing itself.

---

[3] In our own view it is the primary business of a serious phonetics to determine the physical bases of phonological distinctions and not to eke out some kind of justification for the linguist's every phonetic intuition. However, one very recent statement (Malécot, 1970) seems to take the latter point of view, implying that impressionistic phonetic labels are to be taken at face value when naive test subjects can be induced to apply them in conformity with the linguist's own phonetic conviction.

[4] In our view it is irrelevant whether phonetic assertions are conceived to be reflections of physical reality or "part of a theory about the instructions sent from the central nervous system to the speech apparatus" (Postal, 1968:6).

Chomsky and Halle (1968:327-9) have undertaken to account for our data (Lisker and Abramson, 1964:392-413) by supposing that timing variations in the onset of stop voicing result from the interplay of no less than four of their revised set of distinctive features and not from any temporal control of glottal adduction. In fact, at least with respect to segment specification, they seem reluctant to recognize a temporal dimension as an independent feature in their universal set. The four features of the Chomsky-Halle phonetics which together determine voice timing are voice (defined as the state of the larynx appropriate to the generation of voicing or glottal pulsing), tensity, glottal constriction, and heightened subglottal pressure. During stop closure, the feature of tensity, which is defined as a tensing of the supraglottal musculature (so far as consonants are concerned), is supposed to preclude glottal pulsing under otherwise favorable conditions by preventing the pharyngeal expansion said to be necessary for maintenance of an adequate airflow through the glottis.[5] The feature of glottal constriction also functions to prevent pulsing that might otherwise occur during an articulatory closure. Moreover this last feature both allows pulsing to begin promptly with release and, at the same time, inhibits aspiration from developing where it would otherwise arise. In describing how the four features interact, Chomsky and Halle limit them each to two states at the level of phonetic representation: each feature is either present or not present in a given segment.[6]

In the Chomsky-Halle view, six different combinations of values assumed by their four features suffice to explain all the timing relations that we reported in our cross-language study (Lisker and Abramson, 1964). For the languages examined, we had found none in which more than three stop categories could be contrasted with respect to the feature of voice-onset timing. Two of the languages included categories which our measure was clearly incapable of separating, while in a third language there were categories only partly distinguishable on the basis of voicing. Although Chomsky and Halle imply that we did not examine our data as carefully as they have, we did in fact recognize very explicitly that this third language, Korean, is peculiar in having three voiceless categories of initial stop, with slightly and heavily aspirated stops in contrast. We suppose then that the slightly aspirated stop might not be sharply distinguished from the voiceless unaspirated stop solely on the basis of the timing difference and therefore excluded it from consideration when we hazarded the guess that, in most languages of the world, stop categories fall into three phonetic types with respect to voicing time. Chomsky and Halle have elected, however, to recognize four types along this dimension: in one, pulsing begins before release; in a second, it begins immediately upon release; in a third, it lags slightly behind; and in the fourth, considerably behind release. Now, in point of strict phonetic fact, our data can be used to support at least three degrees of voicing lag greater than what we have called "zero lag," particularly if one looks at the timing of stops initial in utterances longer than single words (Figure 1).[7] We might then group together the Korean stop with moderate voicing lag and English /p,t,k/ as a type with first-degree aspiration; the voiceless aspirates in languages such as Cantonese would have second-degree aspiration; and third-degree aspiration would be exemplified by the very strongly aspirated stops found in Korean. Our data would then suggest at least five types of stops occupying different ranges of values along the dimension of voice-onset timing. Since Chomsky and Halle use only six of the twelve allowable combinations of features to explain four timing relations, they might conceivably use certain of the six unused combinations to "handle" additional stop categories. Alternatively, they might invoke the possibility of assigning different scalar values of the features to account for the additional categories. In the present discussion, however, we shall go along with the four stop types as they have described them.

---

[5] Tensity, then, differs somewhat from the "fortis-lenis," or "force of articulation," dimension, whose physical index has generally been taken to be the measure of intraoral air pressure. (See, e.g., Stetson, 1951; Malécot, 1955, 1966b, 1970.) According to the Chomsky-Halle phonetics, the index of tensity would seem to be the absence of laryngeal pulsing during consonant closure when other conditions favor pulsing.

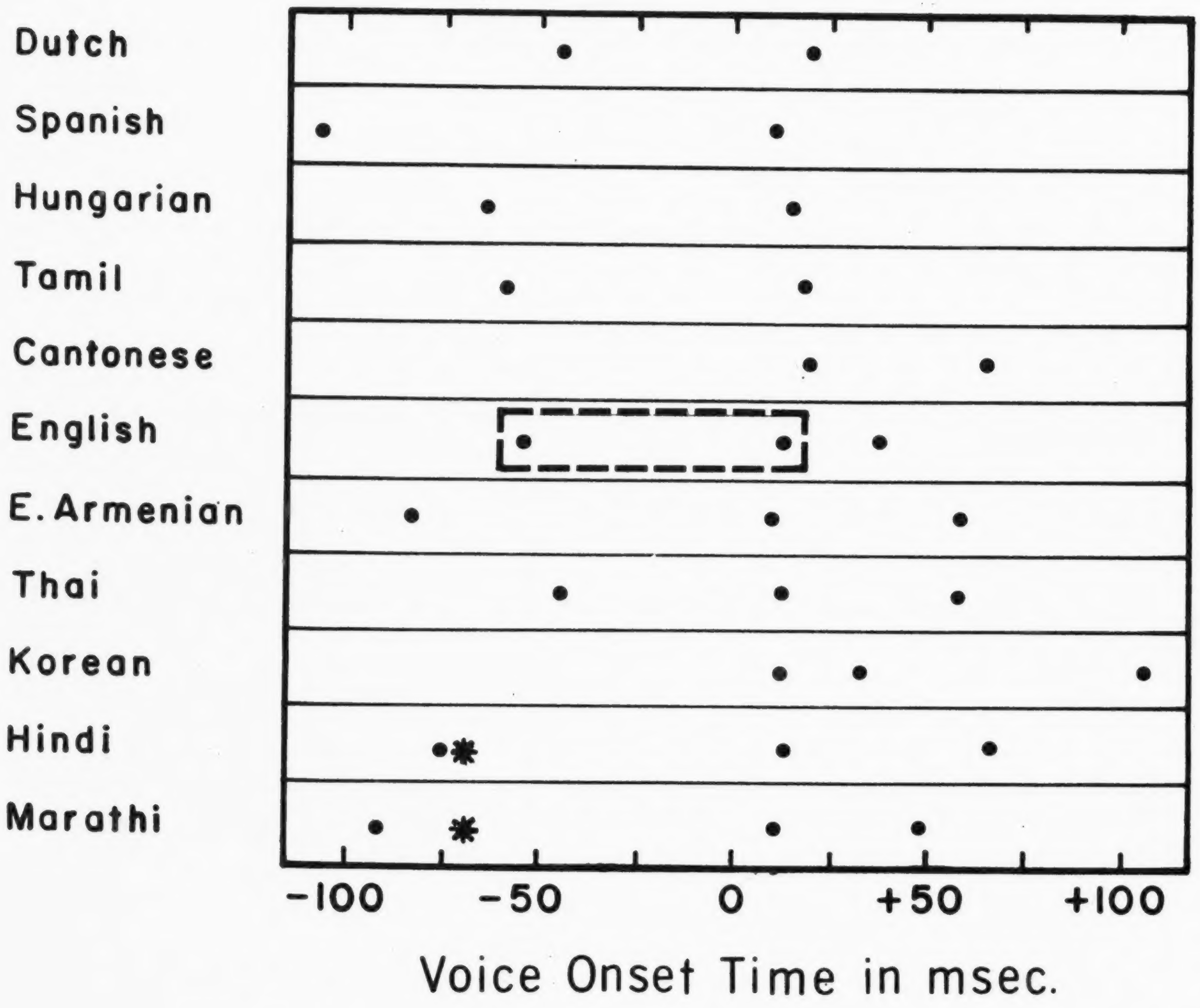
[6] Presumably it is at this point that the Chomsky-Halle model would require that instructions to the speech organs specify how much of each feature is to be used (1968:297-8). The label "yes" in their Table 8 (p. 328) must then be a way of avoiding the problem of assigning scalar values. Given the absence of such data--indeed, the lack of convincing evidence that three of the features are, in fact, generally applicable to languages--this precaution is understandable.

Insofar as their features, if in fact differentially operative in stop production, might affect voice timing, Chomsky and Halle provide perfectly reasonable descriptions of the phonetic consequences of particular combinations of those features. Tensity prevents initiation of pulsing where it would otherwise occur; voice produces pulsing during stop occlusion if there is no tensity and no glottal constriction and otherwise results in onset immediately upon release; heightened subglottal pressure results in the long delay we know as voiceless aspiration if there is neither voice nor glottal constriction, and in the so-called voiced aspiration of Indo-Aryan languages when voice is present and tensity absent. Glottal constriction, as has already been said, prevents both pulsing during closure as well as aspiration, whether voiced or voiceless, where other feature states would favor their development. For the generation of stops with pulsing during closure there is thus voice, no tensity, no glottal constriction, and either an absence of heightened subglottal pressure for the unaspirated or the presence of heightened subglottal pressure for the aspirated voiced stops. For stops characterized by pulsing onset simultaneous with release Chomsky and Halle assume voice to be present, while the lack of closure pulsing is ascribed to tensity and/or a combination of heightened subglottal pressure and glottal constriction. The Korean category of slightly aspirated voiceless stop involves, according to Chomsky and Halle, the absence of all four of their features. The more strongly aspirated Korean stops are produced by tensity and heightened subglottal pressure, in the absence of both voice and glottal constriction. These relations between features and stop category types are summarized in Figure 2, which represents our understanding of their Table 8 (1968:328).

---

[7] Abercrombie (1967:148-9) goes so far as to say that "there can...be many intermediate points...at which voicing sets in: from 'fully voiced' to 'voiceless fully aspirated' is a continuum." Extensive perception testing of one of the present authors (ASA) has yielded five clear labelling categories along our synthetic continuum ranging from 150 msec before stop release to 150 msec after release.

Mean Voice-Onset Times for Stops in Sentence-Initial Position



Stop release is at 0 msec. Starred entries are for voiced aspirates. (From Lisker and Abramson, 1964.)

Fig. 1

Categories of Initial Stops Classified in Terms of Voice-Onset Time and Chomsky-Halle Phonetic Features

Feature	Voicing leads	Voicing coincides	Voicing lags a bit	Voicing lags much
Voice	YES YES	YES YES YES	NO	NO
Tensity	NO NO	NO YES YES	NO	YES
Glottal Constriction	NO NO	NO YES YES	NO	NO
Heightened Subglottal Pressure	NO YES	NO NO YES	NO	YES

Fig. 2

Derived from Chomsky and Halle, 1968:328, Table 8. The broken line shows the same feature complex for two conditions of voice-onset time; this may be an oversight on the authors' part.

Now whatever may be said for the aesthetic appeal and theoretical adequacy, in some abstract sense, of this universal phonetic machinery that Chomsky and Halle have constructed, there remain the serious nonformal questions of its correspondence with well-attested observational data and of the extent to which it simply outruns the data now available. For a few of their suppositions they are able to derive support from certain recent studies. Such evidence, however, is rather skimpier than the tone of flat assertion which Chomsky and Halle adopt would lead the unwary reader to suppose. In two studies of stop consonants in Korean, Kim (1965, 1967) has presented data on voice timing, intraoral air pressure, electromyographic activity, and variations in pharyngeal width, glottal aperture, and the vertical positioning of the larynx. The articulatory information was derived from high-speed X-ray motion pictures of the vocal tract. Information on the articulation of certain of the English stops comes from X-ray measurements by Perkell (1965). These two sets of observations by Kim and Perkell appear to be the sole basis for the Chomsky-Halle description of how the timing of voice onset is controlled in stop production in languages generally; for remarkably enough the bibliographic delving that is manifested in quotations from Winteler (1876) and Sievers (1901) has missed a considerable literature that is both relevant and accessible but does not jibe entirely with the phonetic account they are intent on presenting.

That the Chomsky-Halle mechanism seems complex is in itself no strong argument against it. Complexity of description is required to account for language generally, and as phoneticians we tend to believe that considerations of "economy" are not paramount in determining how speech production is accomplished. Moreover, there can be no quarrel with the view that the larynx does not operate in isolation or that the extralaryngeal components of the Chomsky-Halle mechanism would affect the larynx in the ways they describe, if in fact those components did participate in stop production as they suppose. It is unfortunate, in our view, that Chomsky and Halle have not only been highly selective in what they have chosen to recognize as relevant phonetic observations, but that they have apparently paid only just enough attention to the papers chosen for citation to note those findings which are compatible with their own descriptive scheme. Thus we do not learn from their account of Kim's work, that he concluded from his observations that "it is safe to say now that aspiration is nothing but a function of the glottal opening at the time of release" (Kim, 1967:267), a view not very different from our own feeling that voiceless aspiration is essentially no more than the consequence of delay in the resumption of the voicing position by the larynx (Lisker and Abramson, 1964:416).[8] Nothing in Kim's report or anywhere else in

---

[8] Kim demonstrated, on the basis of X-ray motion pictures, that the different durations of aspiration for the three voiceless stops of Korean correlate directly with different degrees of glottal aperture at the time of release. He supposes (1970:112) that the degree of glottal aperture at release determines how long thereafter the glottis takes to assume the voicing position; this is reasonable if we assume that the rate of glottal closure is relatively constant. Given the present state of our knowledge, however, we insist that there is as yet no solid basis for claiming that these aspects of laryngeal action, size of aperture and voice-onset time, are independently controllable. Nor is it necessarily to be assumed that in utterance-initial position, with which we have mainly been concerned (Lisker and Abramson, 1964), the control must be identical with that exerted in other positions. If we can speak of

the literature provides information as to just how the voiced aspirates of Indo-Aryan languages are produced, nor is it by any means obvious that voiced and voiceless aspiration can be related to one and the same articulatory feature. The postulation of heightened subglottal pressure as the necessary condition for aspiration, both voiced and voiceless, seems at first glance reasonably plausible, but in fact Chomsky and Halle cite no study which establishes a connection between subglottal pressure and any phonetic property associated with individual segments.[9] We might suppose it to be involved in the case of the voiced aspirates, although it seems unlikely that there is no concomitant adjustment of the larynx, but for the voiceless aspirates, at least in English, intraoral air-pressure data [the basis, in fact, for the Chomsky-Halle inferences as to the subglottal situation are Kim's (1965) supraglottal pressure data] strongly suggested that there is no greater pressure than for the unaspirated voiceless stops found in medial posttonic position (Lisker, 1970). More direct evidence showing that subglottal pressure differences between voiced and aspirated voiceless stops in English are negligible has been presented recently by Netsell (1969). As for a relation between tensity and pharyngeal volume, Kim's published records, upon close examination, as often as not show enlargement during the closure of stops said to be tense and no enlargement during the stop without the feature of tensity. Perkell (1969) presents similar data. Moreover, the notion that such pharyngeal enlargement as accompanies the voiced stops and other consonants is a merely passive response to a supraglottal pressure buildup is embarrassed by Perkell's finding a similar enlargement during an English /n/, which may be produced without tensity if one insists but certainly involves no significant pressure buildup to which the observed enlargement could be a passive response. Unless one simply knows in his heart which segments are "tense" and which "lax," it seems just as reasonable to imagine that pharyngeal enlargement is an active adjustment as to see in it confirmation of the absence of a tensity feature. That such an active adjustment is possible has been convincingly argued by Rothenberg (1968) and by Kent and Moll (1969). Of course, in the absence of either active or passive adjustment of the pharyngeal volume, pulsing may be maintained during articulatory closure if the velo-pharyngeal seal is not tight. That this can indeed happen has been shown by Yanagihara and Hyde (1966). Lastly, it must be pointed out that there is no observational basis for considering a feature of glottal constriction that would operate to prevent pulsing under conditions, including the size of glottal aperture, which are otherwise favorable to voicing. Glottal constriction may well be a required feature in any universal phonetics in order to account for phonetic entities like the glottal stop and glottalized consonants, and perhaps also "creaky voice" (Catford, 1964:32), but we have no right at present to suppose that the mechanism by which the vocal folds, and very likely the false cords, are clamped shut can operate to prevent pulsing unless the vocal folds are completely adducted. Moreover, if we would argue on formal rather than substantive grounds, this feature of glottal constriction as a factor controlling voicing-onset time is not only

---

the independent control of degree of glottal aperture and timing of glottal closure, then it would seem to us that Kim is correct in supposing that in noninitial position the speaker controls aperture size; in initial position, on the other hand, it seems to us more reasonable to talk of a control on the timing of glottal closure.

[9] Ladefoged (1967:15-7) alludes to the possibility of such a feature but points to the difficulty of establishing its independent status (p. 87).

highly dubious, but its very necessity depends entirely on the acceptability of two other features of doubtful status, namely tensity and heightened subglottal pressure.

Important to the physical definitions of the Chomsky-Halle distinctive features is the notion of a "neutral speech" configuration of the speech mechanism, one which it is said to assume just prior to the onset of audible activity (1968:300). In the usual case, the "yes" state, i.e., the presence of a feature, represents a greater departure from the neutral position than does the "no." Thus vowels characterized as tense are produced with the body of the tongue further from the neutral position (which is said to be that of the vowel [ɛ]) than it is for otherwise similar vowels which lack this feature. The feature-dimensions high-nonhigh, low-nonlow, coronal-noncoronal, etc., are likewise defined in relation to the posited neutral position of either the body or the blade of the tongue. In the case of the voice feature, however, the situation is somewhat anomolous. The neutral state for voice is defined as that state of the larynx in which glottal pulsing will spontaneously develop when there is a transglottal pressure difference resulting from an unimpeded flow of air through the mouth or nose. The feature of voice is said to be present only for the case where unspecified laryngeal adjustments are assumed necessary to ensure voicing when the supraglottal airway is constricted for stops and fricatives. Under the same condition the absence of the voice feature, on the other hand, entails a large departure from the neutral laryngeal state in that the glottis is sufficiently open to preclude voicing under any condition. The neutral state for the larynx is then compatible with the observation that segments produced with an open oral tract are commonly voiced, while obstruents without contrastive voicing are "normally" voiceless.[10] According to Chomsky and Halle, the development of pulsing during an obstruent constriction requires the presence of the voice feature; during such a constriction the neutral state cannot yield spontaneous vibration of the vocal folds. Thus the absence of pulsing during stop closure ought to be ascribable either to the neutral state or to the absence of the voice feature, but if we read Chomsky and Halle correctly, the neutral voice state is not a permitted one during an occlusion; at any rate it is the voiced state that they ascribe to those voiceless stops for which pulsing begins upon release. Their analysis would require tensity and/or glottal constriction to explain the absence of pulsing during the closure for such stops. The situation would not differ if a neutral glottal state were assumed, for it is not obvious why, if the pharyngeal cavity were free to expand, there would not be a certain amount of closure voicing. Of course, if stops are held to be normally produced without pulsing, then we could suppose that the neutral state of the upper vocal tract is tense. But tensity is also said to characterize vowels for which the body of the tongue takes a position relatively far from the neutral one. Within the phonetic framework provided by Chomsky and Halle the "normal" vowel is scored as neutral in respect to voice and either tense or nontense, while the "normal" stop consonant, it would seem, must be characterized either by the absence of voice or the presence of tensity. The neutral state is said, more or less explicitly, to be nontense in the case of nonobstruents. For the obstruents, on the other hand, either the neutral state is to be considered tense or else it must be characterized as without voice. In either case the notion of a speech-neutral state suffers, for it seems nonsense to talk about a neutral state that

---

[10] See, e.g., Malécot, 1963; Kinkade, 1963; Matina, 1970.

shifts from segment to segment within the utterance, while it seems impossible to define a neutral state that allegedly represents the speech-readiness posture of the vocal tract and at the same time purports to explain why particular constellations of feature states are favored in language generally.[11]

In the case of the other two features which are of interest in connection with stop voicing there is no problem deciding what state is to be equated with the neutral one: both heightened subglottal pressure and glottal constriction must be absent. Finally, it is necessary to point out that the hypothesis of a neutral position of the tract, convenient as it may be for the Chomsky-Halle system of phonetics, rests on only the flimsiest observational basis; certainly there is no solid evidence that the larynx regularly assumes a position just prior to speaking that is independent of the voicing state required for the initial segment of the utterance.

There are other difficulties concealed within the universal phonetics of The Sound Pattern of English. The initial varieties of English /b,d,g/, in their common realizations as stops with pulsing onset at or just after release, are presumably to be taken as voiced, in order to account for the promptness with which pulsing begins; they must, like the voiceless unaspirated stops of Spanish and Korean, be characterized by glottal constriction and, according to the Chomsky-Halle analysis, therefore by tensity as well. Such a representation fails, we believe, to accord with the speaker-linguist whose intuition Chomsky and Halle want to satisfy, or, more importantly, with any available physical evidence. There are data derived from transillumination of the glottis (Lisker et al., 1969) that indicate a closing down of the glottis before the release of initial /b,d,g/, but the time it takes to get from the open breathing position to the onset of pulsing is, according to Lieberman (1967:14), at least 100 msec. It is certainly conceivable that the voicing of these stops begins later than that of the "fully voiced" stops of Spanish, for example, because glottal closure begins earlier in the Spanish case. Another plausible explanation for the absence of closure-voicing in English /b,d,g/ is available if we suppose that pharyngeal enlargement is not simply absence of tensity, but rather a positive gesture that may have little to do with the linguist's intuitive tense-lax dimension. Spanish /b,d,g/ could then be said to involve pharyngeal enlargement as contrasted with English /b,d,g/. Since intraoral air-pressure measurements we have made (Lisker, 1970) indicate no reliable differences between English initial /b,d,g/ and /p,t,k/, either in rate of pressure rise or in peak values, one might reasonably assume that the pharynx is not enlarged for either class of stops. One may still insist on regarding them as nontense and tense, respectively, but only provided one reconsiders the definition of "tensity." Perhaps one should not lightly subsume under one feature the dimensions of pharyngeal size and degree of muscular effort involved in articulatory gestures. As one component of the feature of pharyngeal size one would want to include laryngeal height, since it has been claimed that the larynx is actively lowered during the occlusion of a voiced stop (Stetson, 1951:50, 196-7).

---

[11] We do not here mean to reject the notion of a speech-neutral state out of hand. Rather do we question the plausibility of the Chomsky-Halle statement. In this connection see the recent discussion in Lieberman (1970), particularly his arguments on the need for specifying "language-specific and individual aspects of the neutral state of the vocal tract" (p. 318).

As against Chomsky and Halle's hypothetical picture of how the observed differences in voicing-onset time are generated, we assert the possibility, in the absence of evidence to the contrary, that the speaker exerts some control over the timing of voicing onset by determining the close-down of the glottis. In absolute initial position, the one with which we have been most concerned, it seems not unreasonable to suppose straightforward control of the timing of contraction of certain of the laryngeal muscles. In other positions, however, it appears that the extent of glottal opening, rather than the precise timing of glottal opening and closing, is what is controlled. Evidence for this comes from Kim (1970), in the case of Korean stops, and from our own work on transillumination and fiberoptics photography of the larynx (Lisker et al., 1969, 1970). Moreover we do not mean to assert that differences either in extent or timing of a gesture of glottal opening function in isolation. Certainly air consumption during the release of an aspirated stop is greater than for an unaspirated one (Subtelny et al., 1966; Isshiki and Ringel, 1964; Klatt et al., 1968), and we might expect compensatory adjustments somewhere in the tract. It is not impossible that in producing the voiced aspirates the combination of pulsing with an increase in the rate of airflow through the mouth may be accomplished with the help of an extra pulmonary thrust or that this might also be involved in the production of heavily aspirated voiceless stops. Nor is it unreasonable to expect pharyngeal enlargement during stops with long voicing intervals preceding the release. How consistently these extralaryngeal adjustments are found in running speech, however, is a question that is answerable only on the basis of much more investigation than underlies the Chomsky-Halle phonetic frame. In the absence of such investigation, but with inklings derived from studies currently in progress (Sawashima et al., 1970; Lisker et al., 1970), we prefer to believe that the primary source for the voice timing differences among stop categories is the larynx itself, most particularly in the intrinsic musculature by which degree of glottal opening is regulated.

Like other versions of distinctive feature analysis, the phonetics of Chomsky and Halle implies a more direct concern with the physical tangibles of speech than does the older classificatory system whose basic unit is the segment. The essential purpose of both feature and segment description seems to be pretty much the same: to serve as the basis for a writing system which will enable the linguist to spell any particular expression in any language in a way that incorporates most efficiently, in some sense, his judgments as to how a speaker must manage his vocal tract if he is to produce proper repetitions of that expression. According to Chomsky and Halle it is precisely those judgments that their generative grammar accounts for, with no very precise limits drawn on the explanatory powers of the different components of that grammar. Thus any particular phonetic judgment incorporated in a transcription represents some "mix" of the linguist's semantic, syntactic, and phonetic-phonological knowledge of the specific language. Nor is it excluded that that judgment be informed as to the findings of modern laboratory phonetics. But in view of the announced purpose of Chomsky and Halle in constructing their universal phonetic frame, which is more to explicate the linguist's transcription than to determine rules for generating utterances in the speech mode, it seems fair to say that their phonetic interests are transcriptive rather than descriptive, for there can be little motivation to consider descriptive data that are not reflected in transcriptional practice. This is understandable in that the aim of a linguist's phonetic description is to "capture" speech primarily as the manifestation of some putative digital system, "the language," that underlies it. The digits of the linguist's transcription are the segments, and his phonetic specification of an utterance is tied to the segment in that no more than a single value can be ascribed to each of the features which characterize it. If the digits of the Chomsky-Halle universal phonetics are the segments of the linguist's phonetic transcription, it is important to know exactly what the status of these segments is--whether they reflect a segmentation based on universal phonetic criteria, or whether instead they incorporate knowledge of language-specific phonological traits as well.

In denying that differences in voice-onset timing reflect the speaker's control of the relative timing of laryngeal and articulatory gestures, Chomsky and Halle seem to imply that segments are specifiable as steady states with respect to each of the distinctive features composing them. Thus any articulatory or acoustic shift within a segment is not due to a change in the value assigned to some one or more of its constituent features but is simply the product of their interaction. In the language of present-day syntactic description, such a shift as from the bilabial closure to the aspiration of English initial /p/ would be the surface phonetic effect of a particular combination of a fixed-value features at the deep phonetic level. At this deep level, the one at which control of the phonetic output is effected, changes in the values of features are associated with the shift from one segment to the next.

If the segments of the Chomsky-Halle phonetics have universal validity, we must suppose that the segmentation of a speech stretch can be accomplished independently of any syntactic or semantic knowledge; questions of the type "Is it one segment or two?" simply cannot arise except as there are uncertainties regarding the physical state of affairs. If, on the other hand, segmentation does depend on the linguist's extraphonetic knowledge of the language, and there are grounds for believing this to be the Chomsky-Halle view, then another question must be raised. Let us suppose there exists some stretch of speech which can be uniquely resolved into segments only when we know the phonological rules of the language to which it belongs. Then we might suppose that the speech stretch could equally well be taken to represent sentences in two languages and that the number of segments into which it was analyzed would differ depending on their different phonologies. Such a situation would be, in effect, a case of ambiguity in which two presumably different sequences at some deep phonetic level were identically represented at the surface, i.e., at the level of either articulatory activity or the resulting acoustic signal; such a surface representation is subject to a segmentation based exclusively on the existence of physical discontinuities. It might be argued that, unlike ambiguities of the syntactic-semantic variety, this phonological ambiguity is in principle resolvable in the laboratory, provided the experimental phonetician has access to the deep phonetic facts and can verify just where in the course of production of the utterance there is a change in value for one or more of the phonetic features of the Chomsky-Halle universal set. This, however, would be tantamount to asserting that ultimately there is no possibility of phonological ambiguity, that the segmentation which the linguist practices is uniquely determined by universal phonetic factors. The extraphonetic knowledge applied to this task is thus redundant, however useful it may be to the field linguist without ready access to the deep phonetic level.

If we assume the kind of ambiguity possible where no phonetic criteria are alone sufficient to establish how many segments are needed to specify some utterances, then we are entitled to raise a question of the following kind. Let us suppose, as linguists have, in fact, that an utterance which English speakers would identify as the word pin might be represented phonetically either as [p'ɪn] or [phɪn] and that the choice between the two is dictated by considerations of coding strategy, that is, relative spelling efficiency. If the spelling [p'ɪn] is choosen, as it is if the language is English, then according to Chomsky-Halle the delay in voicing onset results from the interaction of four features having specific values which are fixed for the first segment of this particular representation. If, however, another phonology suggested [phɪn] as the appropriate representation, then presumably both the first and second segments would be characterized by absence of voice. By this second analysis the nonvoice state would be maintained for the duration of two segments rather than one. It appears to us, then, that in denying that speakers exert a temporal control on the larynx, Chomsky and Halle must be referring only to subsegmental control. Control, they seem to be saying, is not of the continuous variety; it can only be applied discretely, in steps the size of their segments. Whether we shall say that the larynx is instructed to maintain the nonvoice state for one or two segments in the case of our ambiguous utterance [p(h)ɪn] depends not entirely on our knowledge of the physical state of affairs with respect to the relevant phonetic features, but on the number of segments we choose to recognize, where the choice is dictated largely by considerations of coding strategy. If a speech stretch such as that preceding the vowel in [p(h)ɪn] may be taken to consist of either one or two segments, depending on considerations of spelling strategy, then it seems difficult to exclude the possibility that, on a purely physical basis, it might be considered to constitute a phonological unit characterized by a delay in voicing onset fully as controlled as would be recognized if the stretch were taken to be composed of two continuous phonological units. To suppose that in the first case the voiceless aspiration is "automatic" and in the second case due to a voicing onset delayed for the duration of one segment has the attractive feature that it provides a link between phonological structure and the speaker's intuition of what the "meaningful" segmentation is. It would at the same time, however, pose a certain threat to the universality of a phonetic theory, if the specification of a speech stretch depended so heavily on knowledge of the language-specific phonological traits that surface-phonetic similarities were obscured.

In our work on stop voicing we have not been concerned with the question of how our measurement data should be related to the phonetic specifications of segments set up according to any particular phonological theory. Instead, taking the word as the object of attention, and in particular those words in whose production there is an initial stoppage of airflow and a subsequent shift to a state of minimal oral obstruction, we have asked simply how far along in the word the larynx begins its audible vibration. The voice-onset time determined for a set of words of the sort just described may serve to characterize distinctively the different categories of stop consonants but might just as well be said to characterize different manners of initiating syllables. In the latter event there is no need to choose between ascribing the feature of interest to the initial stop or to the combination of stop and following vowel. Such a choice is, from the point of view of a physical description, an arbitrary one on the basis of present knowledge. Moreover the problem it poses is artifactual to a mode of description which presumes that the digital mode of representing a linguistic expression in writing must represent the encipherment of an equally discrete sequence of articulatory states assumed in a proper ("ideal") performance of the expression. As linguists, Chomsky and Halle feel obliged to define the task of phonetic specification as one of stating the phonetic properties, not of linguistic expressions but of the segments which they are said to consist of. Their concern is not to determine how an articulatory sequence and its associated acoustic signal, both of them physically neither purely continuous nor purely digital in nature, are related to a linguistic expression but rather to impose digitalization on the physical description in such a way that it will necessarily be a description of the segments in the linguist's spelling of the expression. Chomsky and Halle suppose that a particular combination of fixed values for their phonetic features can generate a sequence of acoustically distinguishable signal elements (e.g., silence + noise burst + noise-excited formant pattern, in the case of a segment [p']), when this sequence as a whole constitutes a single phonological segment. Their supposition may turn out to be true, once we have found ways of collecting relevant measurement data. Pending such empirical confirmation, however, it seems dangerous to us, on another ground, to accept entirely the notion, implied by the Chomsky-Halle reading of our timing data, that the unity of a phonological segment derives from its correspondence to some particular combination of fixed values for their phonetic features. For if a single combination of values may generate a phonological segment decomposable into a sequence of acoustically distinct elements, the possibility cannot be excluded that a single control pattern may activate the speech mechanism so as to produce two or even more phonological segments in a particular sequence. Presumably, in such an event, Chomsky and Halle, and linguists generally, would find unacceptable the notion that such segments must be denied the status of independent phonological elements. But if the objection is raised that the assignment of values to phonetic features has nothing to do with the question of how the control of the vocal tract is managed but rather with the actual state of the vocal tract, then we must recognize that a phonological segment composed of a sequence of acoustically distinct elements reflects just as many states of the vocal tract. Thus we are simply back where we started, with something less than a perfect one-to-one correspondence between phonological segments and units of phonetic description which do not entail recognizing a dimension of continuous temporal control.

Perhaps the questions just raised cannot be answered to the satisfaction of linguists for whom phonetics is primarily the study of speech activity and only secondarily concerned with relating features of that activity to the linguist's transcriptions. Chomsky and Halle have not provided a universal phonetics which describes the speech-producing capabilities of the human vocal tract but instead a phonetics which aims to furnish the linguist with a set of feature values for every symbol of a universal phonetic alphabet. In our study of stop voicing we wanted to determine how effective the measure of relative voicing-onset time was as a basis for distinguishing physically among homorganic stop categories, and our findings suggested that it might very well be more effective than any other single physical measure. We were also interested in evidence that would lead us to suppose that the timing of voicing onset is subject to constraints severe enough to mean that this dimension does not constitute an articulatory continuum. Our data, while they suggested that certain values of voice-onset time are preferred generally by speakers, did not provide strong support for rejecting the possibility that speakers are capable of producing stops with any duration of voicing lead or lag over a range of several hundred milliseconds. Chomsky and Halle, on the other hand, starting with the notion of the phonological segment defined as a set of features with fixed values, have found no compelling reason to admit the possibility of a continuous control of voice-onset timing. It might be said that these two viewpoints are not really opposed, in the sense that only one at most can be correct; together they simply represent a reiteration of the old well-known "segmentation problem," in that both may be correct, or at least useful approximations of the truth, for the different interests they represent. This view seems to be implied when Chomsky and Halle state that "since the phonetic transcription...represents the speaker-hearer's interpretation rather than directly observable properties of the signal...there is no longer a problem that the transcription is composed of discrete symbols whereas the signal is quasi-continuous...." (1968:294). If Chomsky and Halle had claimed for their universal phonetics only that it adequately incorporated the directly observable properties of the vocal tract during speech production insofar as these can be accommodated to the segmental mode of description adopted for structural linguistic reasons, then the motivation for their excluding from consideration the possibility of a continuous control of timing would have been clear. Instead they chose to assert that, as a matter of physical fact, the speaker does not have the capacity to exert control over voice-onset timing. That they mean to make a claim of a substantive rather than a merely formal nature is indicated, moreover, by the fact that, in another context, they feel obliged to say that "phonetically we have to recognize a feature that governs the timing of different movements within the limits of a single segment" (1968:317). If the features of tensity, heightened subglottal pressure, and glottal constriction are "at our disposal" (p. 327) as features which serve to control voice-onset timing because on other grounds it seems necessary to consider them members of the universal set of features, then by the same peculiar kind of reasoning a feature of intrasegmental timing should be "available" for describing the differences in voice-onset timing between stop categories. If some criterion of economy, intuitively reasonable in establishing the elements of word and sentence structures, is applied in the formulation of phonetic description, with the result that the marshalling of phonetic features becomes a *tour de force* which outruns present knowledge and contravenes available data, then this criterion must be rejected. With it goes the last reason for accepting the Chomsky-Halle analysis in preference to a straightforward account of stop consonant distinctions in terms of laryngeal timing control.

REFERENCES

Abercrombie, D. (1967) *Elements of General Phonetics*. (Chicago: Aldine).

Abramson, A.S. and L. Lisker (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In *Proc. Fifth Intl. Cong. on Acoustics*, ed. by D.E. Commins, A51. (Liège: Imp. G. Thone).

Abramson, A.S. and L. Lisker. (1970) Discriminability along the voicing continuum: Cross-language tests. In *Proc. Sixth Intl. Cong. Phon. Sci., Prague 1967*, 569-73. (Prague: Academia).

Bloomfield, L. (1933) *Language*. (New York: Holt).

Catford, J.C. (1964) Phonation types: The classification of some laryngeal components of speech production. In *In Honour of Daniel Jones*, ed. by D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott and J.L.M. Trim, 26-37. (London: Longmans).

Chomsky, N. and M. Halle. (1968) *The Sound Pattern of English*. (New York: Harper and Row).

Fromkin, V. (1966) Neuromuscular specification of linguistic units. Language and Speech *9*, 170-99.

Harris, K.S., G.F. Lysaught, and M.H. Schvey. (1965) Some aspects of the production of oral and nasal labial stops. Language and Speech *8*, 135-47.

Haugen, E. (1951) Directions in modern linguistics. Lg. *27*, 211-22.

Isshiki, N. and R. Ringel. (1964) Airflow during the production of selected consonants. JSHR *7*, 233-44.

Kent, R.D. and K.L. Moll. (1969) Vocal-tract characteristics of the stop cognates. J. acoust. Soc. Am. 46, 1549-55.

Kim, C.-W. (1965) On the autonomy of the tensity feature in stop classification (with special reference to Korean stops). Word 21, 339-59.

Kim, C.-W. (1967) A cineradiographic study of Korean stops and a note on "aspiration." Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., 86, 259-71.

Kim, C.-W. (1970) A theory of aspiration. Phonetica 21, 107-16.

Kinkade, M.D. (1963) Phonology and morphology of upper Chehalis: I. IJAL 29, 181-95.

Klatt, D.H., K.N. Stevens, and J. Mead. (1968) Studies of articulatory activity and airflow during speech. In Sound Production in Man, ed. by Arend Bouhuys, 42-55, (Annals of the New York Academy of Sciences, 155, Art. 1.).

Ladefoged, P. (1967) Linguistic phonetics. University of California, Los Angeles, Working Papers in Phonetics, 6.

Lieberman, P. (1967) Intonation, perception, and language. Research Monograph 38 (Cambridge: M.I.T. Press).

Lieberman, P. (1970) Towards a unified phonetic theory. Linguistic Inquiry 1, 307-32.

Lisker, L. (1970) Supraglottal air pressure in the production of English stops. Language and Speech 13, 215-30.

Lisker, L., F.S. Cooper, and A.M. Liberman. (1962) The uses of experiment in language description. Word 18, 82-106.

Lisker, L. and A.S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Lisker, L. and A.S. Abramson. (1967) Some effects of context on voice onset time in English stops. Language and Speech 10, 1-28.

Lisker, L., A.S. Abramson, F.S. Cooper, and M.H. Schvey. (1969) Transillumination of the larynx in running speech. J. acoust. Soc. Am. 45, 1544-46.

Lisker, L., M. Sawashima, A.S. Abramson, and F.S. Cooper. (1970) Cinegraphic observations of the larynx during voiced and voiceless stops. J. acoust. Soc. Am. 48, 119.

Lubker, J.F. and P.J. Parris. (1970) Simultaneous measurements of intraoral pressure, force of labial contact, and labial electromyographic activity during production of the stop consonant cognates /p/ and /b/. J. acoust. Soc. Am. 47, 625-33.

Malécot, A. (1955) An experimental study of force of articulation. Studia Linguistica 9, 35-44.

Malécot, A. (1963) Luiseño, a structural analysis I: Phonology. IJAL 29, 89-95.

Malécot, A. (1966a) Mechanical pressure as an index of "force of articulation." Phonetica 14, 168-80.

Malécot, A. (1966b) The effectiveness of intra-oral air-pressure-pulse parameters in distinguishing between stop cognates. Phonetica 14, 65-81.

Malécot, A. (1970) The lenis-fortis opposition: Its physiological parameters. J. acoust. Soc. Am. 47, 1588-92.

Mattina, A. (1970) Phonology of Alaskan Eskimo, Kuskokwim dialect. IJAL 36, 38-45.

Mattingly, I.G. and A.M. Liberman. (1969) The speech code and the physiology of language. In Information Processing in the Nervous System, ed. by K.N. Leibovic, 97-117 (New York: Springer).

Netsell, R. (1969) Subglottal and intraoral air pressures during the intervocalic contrast of /t/ and /d/. Phonetica 20, 68-73.

Perkell, J.S.(1965) Cineradiographic studies of speech: Implications of a detailed analysis of certain articulatory movements. In *Proc. Fifth Intl. Cong. on Acoustics*, ed. by D.E. Commins, A32 (Liège: Imp. G. Thone).

Perkell, J.S. (1969) *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. (Cambridge: M.I.T. Press).

Postal, P.M.(1968) *Aspects of Phonological Theory*. (New York: Harper and Row).

Rothenberg, M. (1968) The breath-stream dynamics of simple-released-plosive production. Bibliotheca Phonetica *6*. (Basel: S. Karger.)

Sawashima, M., A.S. Abramson, F.S. Cooper, and L. Lisker. (1970) Observing laryngeal adjustments during running speech by use of a fiberoptics system. Phonetica *22*, 193-201.

Sievers, E. (1901) *Grundzüge der Phonetik*. (Leipzig: Breitkopf and Härtel).

Stetson, R.H. (1951) *Motor Phonetics*, 2nd ed. (Amsterdam: North Holland).

Subtelny, J.D., J.H. Worth, and M. Sakuda. (1966) Intraoral pressure and rate of flow during speech. JSHR *9*, 498-518.

Tatham, M.A.A. and K. Morton. (1969) Some electromyographic data towards a model of speech production. Language and Speech *12*, 39-53.

Winteler, J.C. (1876) *Die Kerenzer Mundart des Kantons Glarus in ihren Grundzügen Dargestellt*. (Heidelberg: Carl Winter).

Yanagihara, N. and C. Hyde. (1966) An aerodynamic study of the articulatory mechanism in the production of bilabial stop consonants. Studia Phonologica *4*, 70-80.

Auditory and Linguistic Processes in the Perception of Intonation Contours*

Michael Studdert-Kennedy+ and Kerstin Hadding++
Haskins Laboratories, New Haven

ABSTRACT

The fundamental frequency contour of a 700-msec vocoded utterance, "November" [no'vɛmbɚ], was systematically varied to produce 72 contours, different in f_o at the stress and over the terminal glide. The contours were recorded (1) carried on the speech wave, (2) as modulated sine waves. Swedish and American subjects classified (1) both speech and sine-wave contours as either terminally rising or terminally falling (psychophysical judgments), (2) speech contours as questions or statements (linguistic judgments). For both groups, two factors acted in complementary relation to govern linguistic judgments: perceived terminal glide and f_o at the stress. Listeners tended to classify contours with an apparent terminal rise and/or high stress as questions, contours with an apparent terminal fall and/or low stress as statements. For both speech and sine waves psychophysical judgments of terminal glide were influenced by earlier sections of the contour, but the effects were reduced for sine-wave contours, and there were several instances in which speech psychophysical judgments followed the linguistic more closely than the sine-wave judgments. It is suggested that these instances may reflect the control exerted by linguistic decision over perceived auditory shape.

The perception of spoken language may be conceived as a process conducted at several successive and simultaneous levels. Auditory, phonetic, phonological, syntactic, and semantic processes form a hierarchy, but decisions from higher levels also feed back to correct or verify tentative decisions at lower levels and to construct the final percept. Suitable experiments (e.g., Warren, 1970) may demonstrate the control exercised by higher on lower level decisions, and the partial determination of phonetic shape by phonological and syntactic rules is readily assumed by some linguists (e.g., Chomsky and Halle, 1968, p. 24). However, the auditory level, itself a complex of interactive processes by which an acoustic signal is converted into a representation suitable for input to the phonetic component (Fourcin, in press), is commonly taken to be relatively independent.

A few studies have questioned this assumption. Ladefoged and McKinney (1963), for example, showed that judgments of the loudness of words presented in a carrier sentence may be more closely related to the work done upon them in phonation, that is, to their degree of stress, than to their acoustic intensity. Allen (1971), replicating and extending the experiment, showed that both acoustic level and inferred vocal effort may serve as cues for the loudness of speech, and that individuals differ in the weight they assign to these cues. Evidently, loudness judgment of speech may entail a relatively complex process of inference, drawing upon more than one level of analysis. The same may be true of pitch judgment: Hadding-Koch and Studdert-Kennedy (1963, 1964, 1965) found that auditory judgments of listeners, asked to assess fundamental frequency (f_o) contours imposed synthetically on a carrier word, seemed to be influenced by linguistic decisions. The present experiment extends this earlier work and, by examining the relations among sections of the f_o contour used in judging an utterance as a question or statement, attempts a more detailed understanding of auditory-linguistic interaction in the perception of intonation countours.[1]

---

*The results of this study were reported at the Seventh International Congress of Phonetic Sciences, Montreal, Canada, August 1971 and will be published in the Proceedings.

+Also Graduate Center and Queens College, City University of New York.

++Visiting Research Associate from Lund University, Sweden.

The starting point for the study is the importance commonly attributed to the terminal glide as an acoustic cue for judgment of an utterance as a question or statement. Two related sets of questions present themselves. The first concerns the basis for auditory judgments of the glide. From our earlier study (Hadding-Koch and Studdert-Kennedy, 1963, 1964, 1965) it was evident that listeners frequently judge a falling glide as rising and a rising glide as falling. Is the origin of this effect auditory (psychophysical) or linguistic? Our study left the question unanswered. There, we systematically manipulated the contour of an utterance by varying f_o at the stress peak, at the "turning point" before the terminal glide, and at the end point We then asked listeners to classify each contour as (1) question or statement (linguistic judgment), (2) having a terminal rise or fall (psychophysical judgment). The two tasks yielded remarkably similar results: whether judging the entire contour linguistically or its terminal glide psychophysically, listeners were influenced in similar ways by the overall pattern of the contour. The outcome suggested that auditory judgments may have been controlled, in part, by linguistic judgments. But the reverse interpretation--that linguistic judgments of the entire contour were controlled by auditory judgments of the terminal glide--is equally plausible as long as we do not know the auditory capacity of listeners for judging the terminal glides of matched nonspeech contours. The present study attempts to resolve this ambiguity by including the necessary nonspeech judgments. Effects observed only in the two types of speech judgment would then be compatible with the first interpretation, while effects observed in all three types of judgment would be compatible with the second.

At the same time, this study broaches a second, related set of questions. These concern the roles of the various sections of the contour in determining linguistic judgments. Previous studies, both naturalistic and experimental, have suggested that listeners make use of an entire contour, not simply of the terminal glide, in judging an utterance (see Gårding and Abramson, 1965; Hadding-Koch, 1961; Hadding-Koch and Studdert-Kennedy, 1963, 1964, 1965). For example, spectrographic analyses of Swedish speech have shown that, in this language, "yes-no" questions normally display not only a terminal rise, but also an overall higher f_o than statements (Hadding-Koch, 1961). Other utterances in which the speaker wants to draw the listener's special attention also display an overall high f_o and a terminal rise: in listening tests the labels "question," "surprise," "interest" have been found to be interchangeable (Hadding-Koch, 1961, pp. 126 ff.). If a speaker is not interested or is asking a question to which he thinks he knows the answer,[2] his utterances tend to display a lower overall f_o and a falling terminal glide, similar to those of statements.

---

[1] The acoustic correlates of intonation are said to be changes in one or more of three variables: fundamental frequency, intensity, and duration, with variations in fundamental frequency over time being the strongest single cue (Bolinger, 1958; Denes, 1959; Fry, 1968; Lehiste, 1970; Lieberman, in press). The present study is concerned with only one of these variables, fundamental frequency, and the term "intonation contour" refers exclusively to contours of fundamental frequency.

The importance of the entire contour may be reflected in the phonetic description. If four f_o levels are postulated, with arrows showing the direction of the terminal glide, the intonation contour of a typical Swedish "yes-no" question could be described with one number at the beginning of the utterance and two at the stress,[3] as $3\ 44\ 2{\uparrow}^{3}$ (the superscript 3 indicates the end point of the terminal glide) or, if less "interested," as $2\ 33\ 2{\uparrow}^{3}$. A neutral statement would be best described as 2 33 1↓, or even 2 22 1↓, though the latter might also indicate a certain indifference. Much the same statement contour is typical of American English. However, questions in this language are said to display a more or less continuously rising contour (Pike, 1945; Hockett, 1955) which might be described as $2\ 22\ 3{\uparrow}^{4}$ or $2\ 33\ 3{\uparrow}^{4}$. Similar contours occur in Swedish echo-questions.[4]

These naturalistic observations of speech are, in general, consistent with results of our experimental study of perception (Hadding-Koch and Studdert-Kennedy, 1963, 1964, 1965). Swedish listeners selected a typical Swedish question (2 44 2↑) among their preferred question contours, and a lower contour with a level terminal glide (2 33 1→) among their preferred statements (they would probably have preferred 2 33 1↓ for a statement had this contour been included). The North American listeners also preferred 2 44 2↑ for a question and 2 33 1→ for a statement, but they were more uncertain (in less agreement with one another) than the Swedish listeners--perhaps because the contours were based on Swedish speech and did not include, for example, a typical American English question.

---

[2] Many workers who have reported, for various languages, that the same intonation is used in questions as in statements, seem to have been anxious to exclude all emotional "overtones" and therefore told their subjects to speak in a neutral voice. The result is that, in the absence of grammatical Q-markers, utterances sound like statements. A "neutral" intonation is not enough to convey, as sole cue, the impression of a question. If a question is asked merely for form's sake, with no particular interest in the answer, no difference in intonation is to be expected from that of a statement.

[3] We write two numerals at the stress and one at the turning point, even though they may be on the same "level" (intonation level, f_o level), cf. Delattre, 1963; Hockett, 1955.

[4] Compare the similar difference in intonation contours for French suggested by Léon, in press.

Granted, then, the importance of the entire contour, we may now ask how its various sections work together to control linguistic judgment. Here, let us recall a central finding of our previous study, namely that there was perceptual reciprocity among various sections of a contour: listeners would trade a high f_o at one point in the utterance for a high f_o elsewhere. For example, an utterance with a relatively high f_o at peak or turning point required a smaller terminal rise to be heard as a question than an utterance with relatively low f_o at peak or turning point. We may interpret this reciprocity in either of two ways. The first interpretation assigns only auditory status to peak and turning point and assumes their linguistic role to be indirect. Thus, an utterance is marked as question or statement by its apparent terminal glide. Earlier sections of the contour are important only insofar as they alter (by some mechanism to be specified) listeners' perceptions of that glide and thereby give rise to the observed reciprocity effects. Lieberman's (1967) account of our results rests squarely on these assumptions. He selects an "analysis-by-synthesis" mechanism to account for the reciprocity.

An alternative interpretation assigns a direct linguistic function to peak and turning point. An utterance is marked as question or statement not only by its terminal glide, but also by the f_o pattern over its earlier course. Listeners discover at least two acoustic cues within a contour, either or both of which may control their linguistic decision. The weighting of these cues (by some unknown mechanism) gives rise to the reciprocity observed in linguistic judgments.

A second purpose of this study was to distinguish between these accounts, again by extending our earlier work to include judgments of the terminal glides of matched nonspeech contours. Effects present in all three types of judgment would then require the first interpretation but would exclude an account, such as that of Lieberman (1967), that invoked specialized speech mechanisms. Effects present only in the two types of speech judgment would be compatible with both the first interpretation and Lieberman's hypothesized mechanism. Effects present only in the linguistic judgments would require the second interpretation.

Finally, an additional purpose of the study was to extend our cross-linguistic comparison of Swedish and American English listeners. We therefore enlarged the set of contours to include typical questions and statements from both American English and Swedish.

METHOD

The stimuli were prepared by means of the Haskins Laboratories Digital Spectrum Manipulator (DSM) (Cooper, 1965). This device provides a spectrographic display of a 19-channel vocoder analysis, digitized to 6 bits at 10-msec intervals, and permits the experimenter to vary the contents of each cell in the frequency-time matrix, before resynthesis by the vocoder. For the present study we were interested in the channel that displayed the time course of the fundamental frequency of the utterance, since it was by manipulating the contents of this channel that we varied f_o.

The utterance "November" [no'vɛmbɚ] was spoken by an American male voice into the vocoder and stored in the DSM. F_o was then manipulated over a range from 85 cps to 220 cps. The f_o values at the most important points of the contours (starting point, peak, turning point, and end point) were chosen to represent four different f_o levels of a speaker with a range from 65 cps to 250 cps. The four levels were based on a previous analysis of a long sample of speech by a speaker with this particular range (Hadding-Koch, 1961, p. 110 ff.).[5]

The contours are schematized in Figure 1. They range between two poles that may be marked 2 44 3↑4 and 2 11 1↓. All contours start on a f_o of 130 Hz (level 2), sustained for 170 msec, over the first syllable. They then move, during 106 msec, to one of three peaks: 130 Hz (L, or low, level 2), 160 Hz (H, or high, level 3), 200 Hz (S, or superhigh, level 4). They proceed, during 127 msec, to one of four turning points: 100 Hz (high level 1), 120 Hz (level 2), 145 Hz (low level 3), 180 Hz (high level 3). Finally, they proceed, during 201 msec, to one of six end-points: 85 Hz (level 1), 100 Hz (high level 1), 120 Hz, 145 Hz, 180 Hz, and 220 Hz (level 4). Peak, turning point, and end point are each sustained for 32 msec. The combination of three peaks, four turning points, and six end points yields 72 contours, each specified by a letter and two numbers (e.g., S24, L36) and each lasting 700 msec.

The 72 contours were recorded on magnetic tape from the output of the vocoder in three forms: (1) carried on a speech wave [no'vɛmbɚ], (2) as a frequency-modulated sine wave, (3) as a frequency-modulated train of pulses. Each set of 72 was spliced into five different random orders with a five-second interval between stimuli and a ten-second pause after every tenth stimulus. They were presented to Swedish and U.S. subjects as described below.

Swedish Subjects. Twenty-two graduate and undergraduate volunteers were tested in three sessions, each lasting about 45 minutes. They listened to the tests over a loud speaker at a comfortable listening level in a quite room. In a given session they heard the five test orders for one type of stimulus

---

[5] One of the contentions of that study, based on a number of utterances in continuous speech by several Swedish subjects, was that every speaker has, in addition to a general speaking range, clusters of "favorite pitches" which he uses, for instance, on stressed segments of statements (represented by the H-peak in the present study), and a higher level which he uses for questions and various expressions of "interest" (here represented by level 4; cf. also Bolinger, 1964).

Statements were found in that study to end on a low level, hesitant or exclamatory utterances, higher up. Questions tended to have a terminal rise, usually from level 2, or a fall ending comparatively high. Questions were also generally spoken with an overall high f_o compared to statements, a phenomenon that, according to the literature, occurs in many languages (Hermann, 1942; Bolinger, 1964). The contour then often started high. Polite or friendly statements too might end with a final rise, but from a comparatively low level and with a moderate range (cf. Uldall, 1962).

Schema of Fundamental Frequency Contours
Imposed on the Utterance "November" [noˈvɛmbɚ]

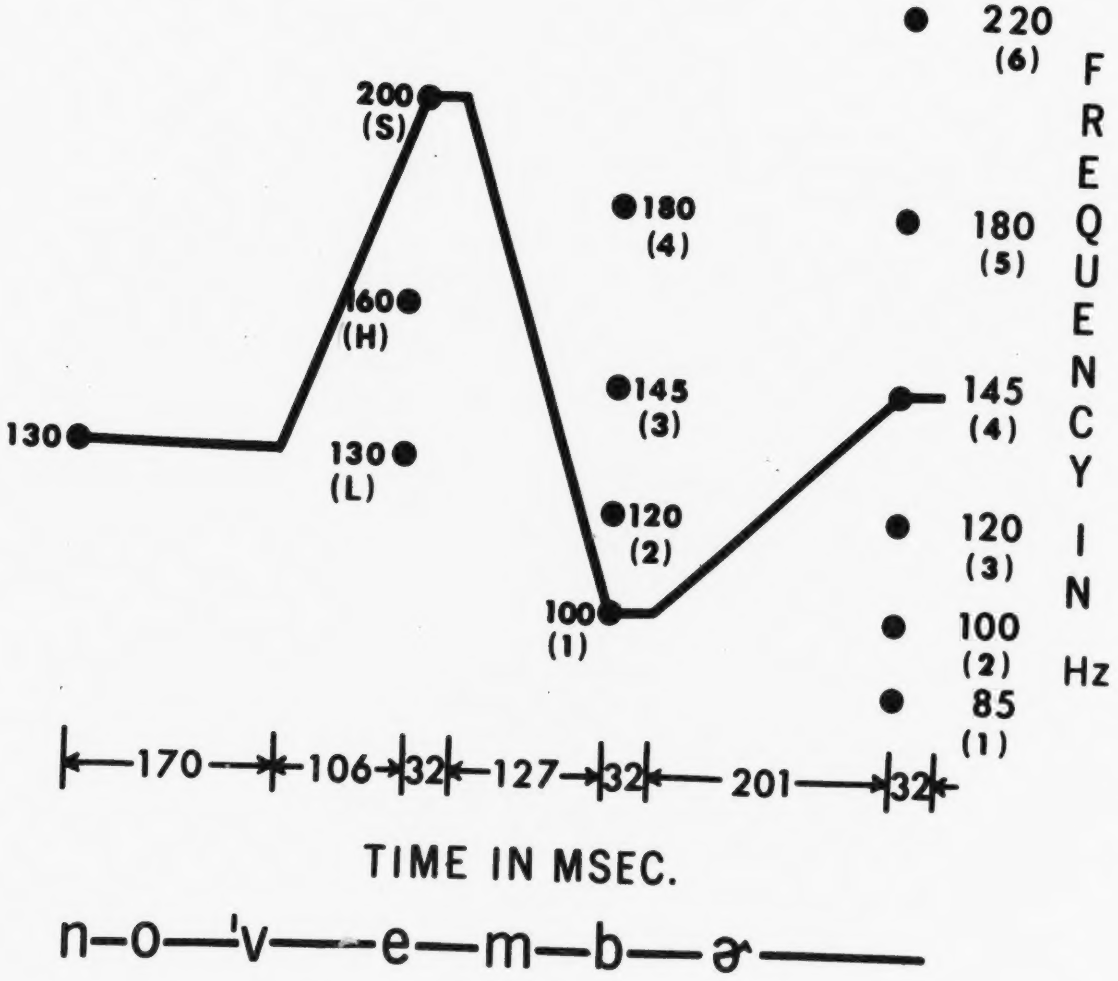


Fig. I

only. They were divided into two groups of 11. Both groups heard the sine-wave stimuli first; this was an important precaution intended to exclude any possible influence of speech mechanisms on judgments of the nonspeech stimuli. In the second and third sessions both groups made psychophysical or linguistic judgments on the speech stimuli, group 1 in the order psychophysical-linguistic, group 2 in the reverse order. In the sine-wave session and in the psychophysical speech session, subjects were asked to listen to the final glide of each contour and judge whether it was rising or falling. In the linguistic speech session subjects were asked to judge each contour as more like a question or more like a statement. For each contour, the procedure yielded 5 judgments by each subject under each condition, a total of 110 judgments in all.

U.S. Subjects. Sixteen female undergraduate paid volunteers were divided into two groups of eight. The procedure duplicated that followed with the Swedish subjects, except that the U.S. subjects listened to the tests over earphones in individual booths. The output of the phones was adjusted by means of a calibration tone to be approximately 75 db SPL. These subjects also made psychophysical judgments on the pulse-train stimuli; these were counterbalanced with the sine waves in the first two sessions before the speech stimuli had been heard. The procedure yielded a total of 80 judgments on each contour under each condition.

RESULTS

No systematic differences between groups due to the order in which they made their judgments were observed. Data are therefore presented for the combined groups throughout. Figures 2 and 4 display the Swedish data, Figures 3 and 5, the U.S. data. In each figure the left column gives the linguistic, the middle column the speech psychophysical, and the right column the sine-wave data.[6] Percentages of question and statement judgments (linguistic) or of rise and fall judgments (speech psychophysical and sine-wave) are plotted against terminal glide, measured as rise (positive) or fall (negative) in Hz, from turning point to end point. In Figures 2 and 3 parameters of the curves are f_o values at peaks (S, H, L), displayed for the four turning-point f_o values from 1 (top) to 4 (bottom). In Figures 4 and 5 parameters of the curves are f_o values at turning points (1, 2, 3, 4) displayed for the three peak f_o values of S(top), H (middle), and L (bottom).

Linguistic Judgments

Cross-Language Comparisons

Before considering the acoustic variables controlling linguistic judgments, we will briefly compare Swedish and U.S. results. The main drift of the data is very similar for the two groups. A broad description of preferred statement and question contours for both groups can be given.

Statements. Figure 6 schematizes the most frequently preferred contours, those obtaining 90% or better agreement. For all these contours, except two (L13; H13, Swedish only), the final f_o of the terminal glide is the lowest f_o

---

[6] Judgments of the modulated sine waves and pulse trains by U.S. subjects were essentially identical. Accordingly, only sine-wave data are presented here.

Percentages of Question or Rise Responses (left-axis)
and Statement or Fall Responses (right-axis)
Plotted as Functions of Terminal Glide in Hz

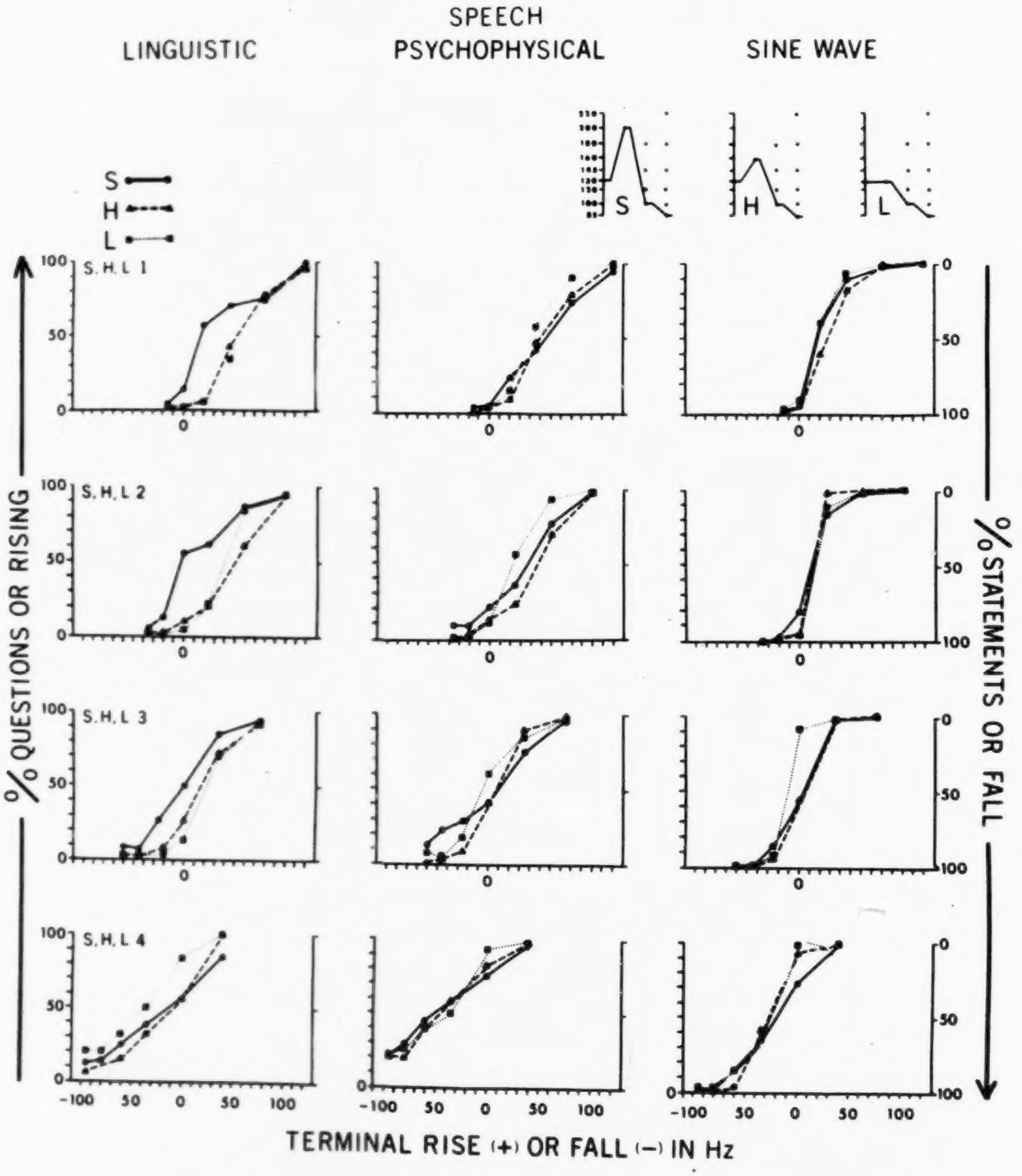


Fig. 2

Turning-point values are constant across rows and peak values are parameters of the curves. For Swedish subjects.

Percentages of Question or Rise Responses (left-axis)
and Statement or Fall Responses (right-axis)
Plotted as Functions of Terminal Glide in Hz

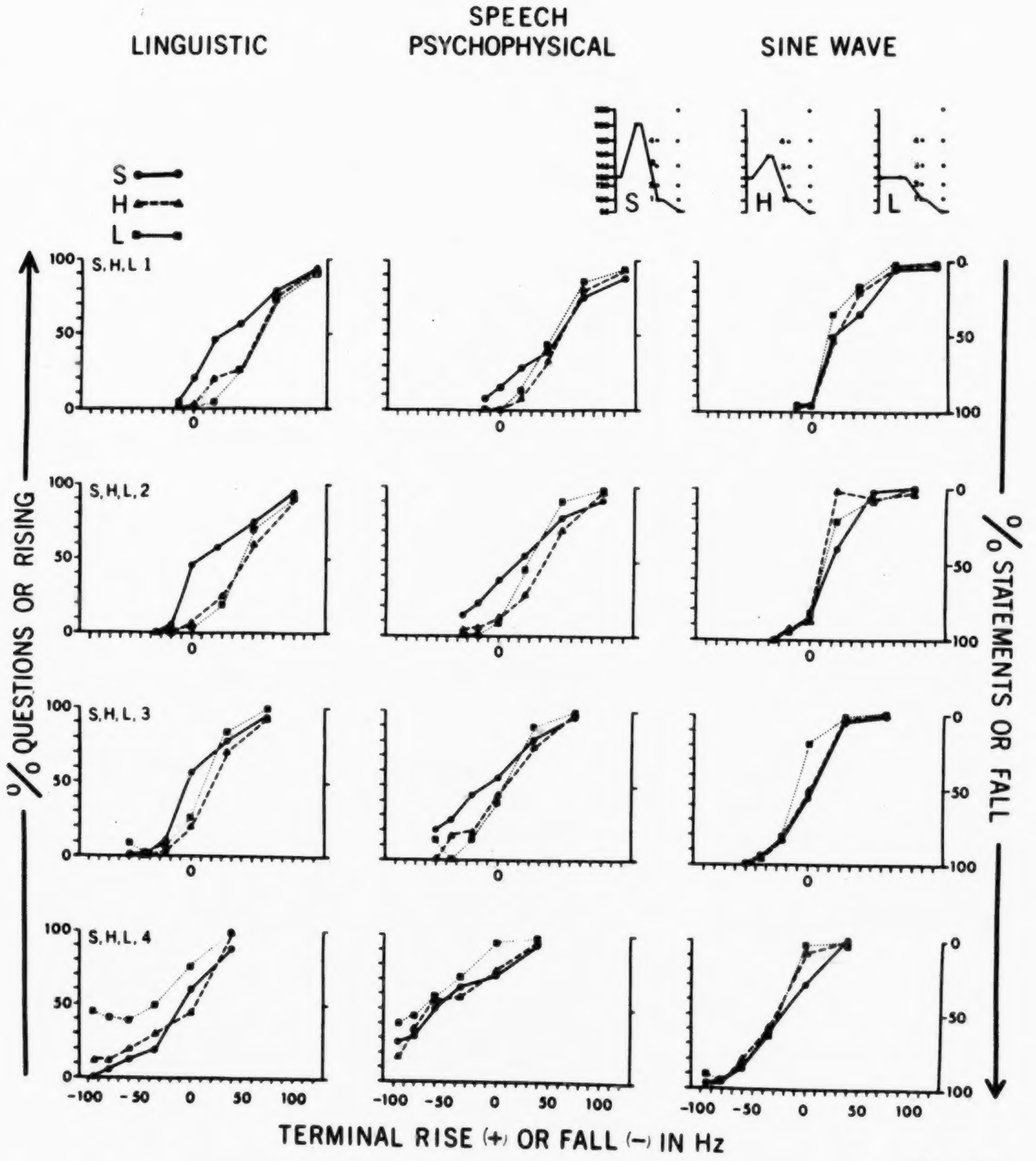
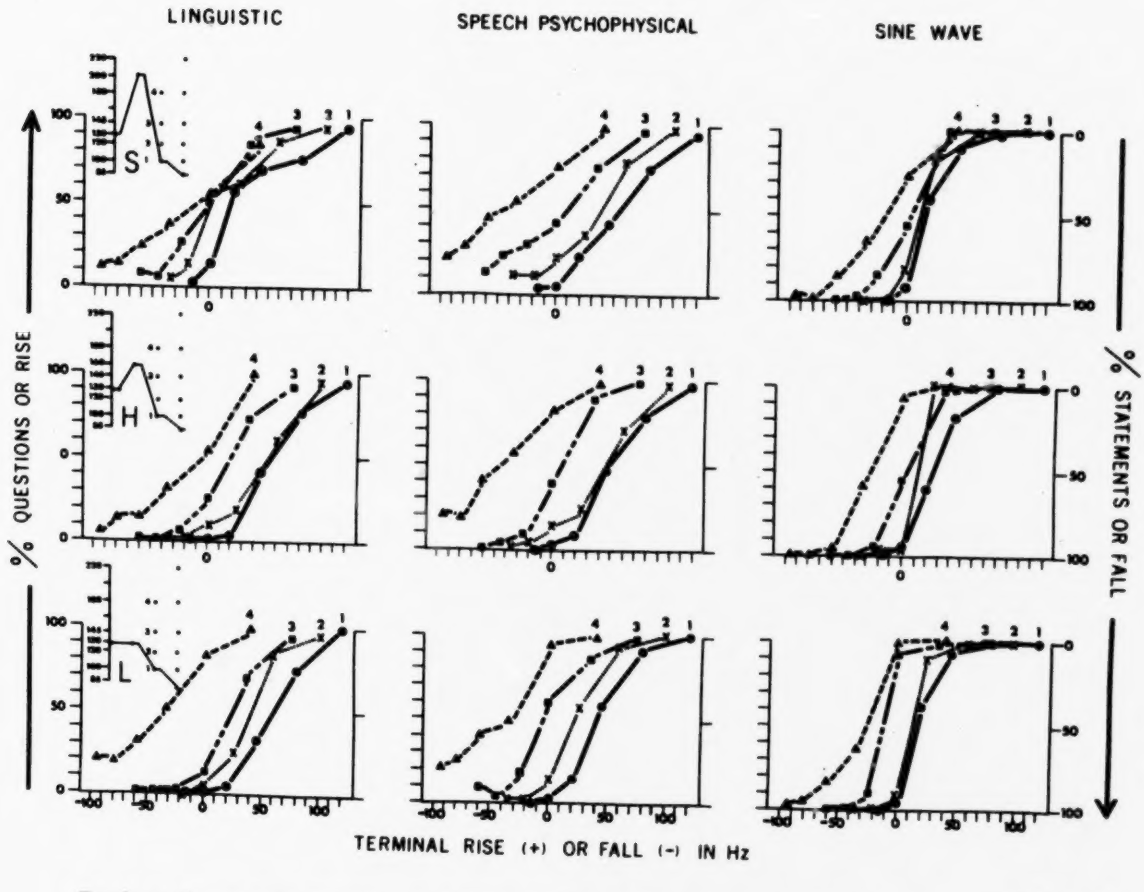


Fig. 3

Turning-point values are constant across rows and peak values are parameters of the curves. For American subjects.

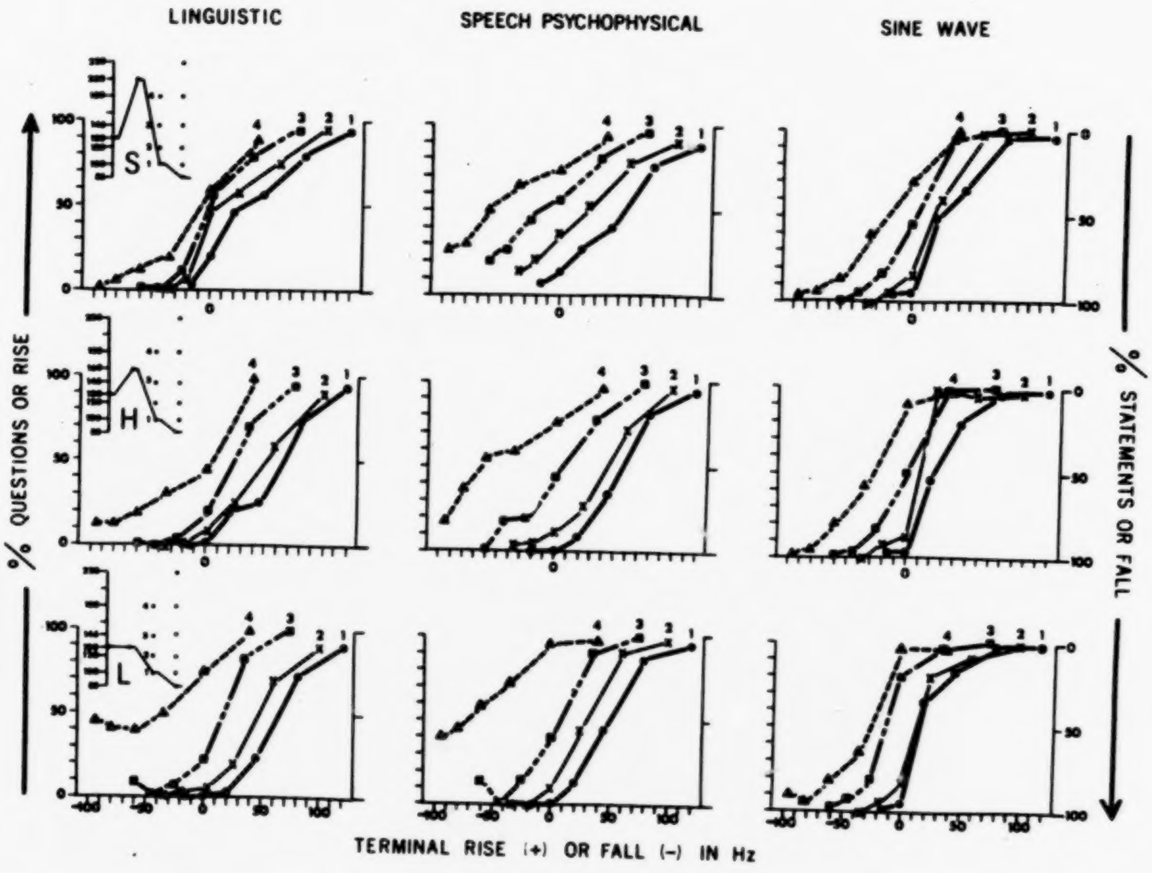
Percentages of Question or Rise Responses (left-axis)
and Statement or Fall Responses (right-axis)
Plotted as Functions of Terminal Glide in Hz



Peak values are constant across rows and turning points are parameters of the curves. For Swedish subjects.

Fig. 4

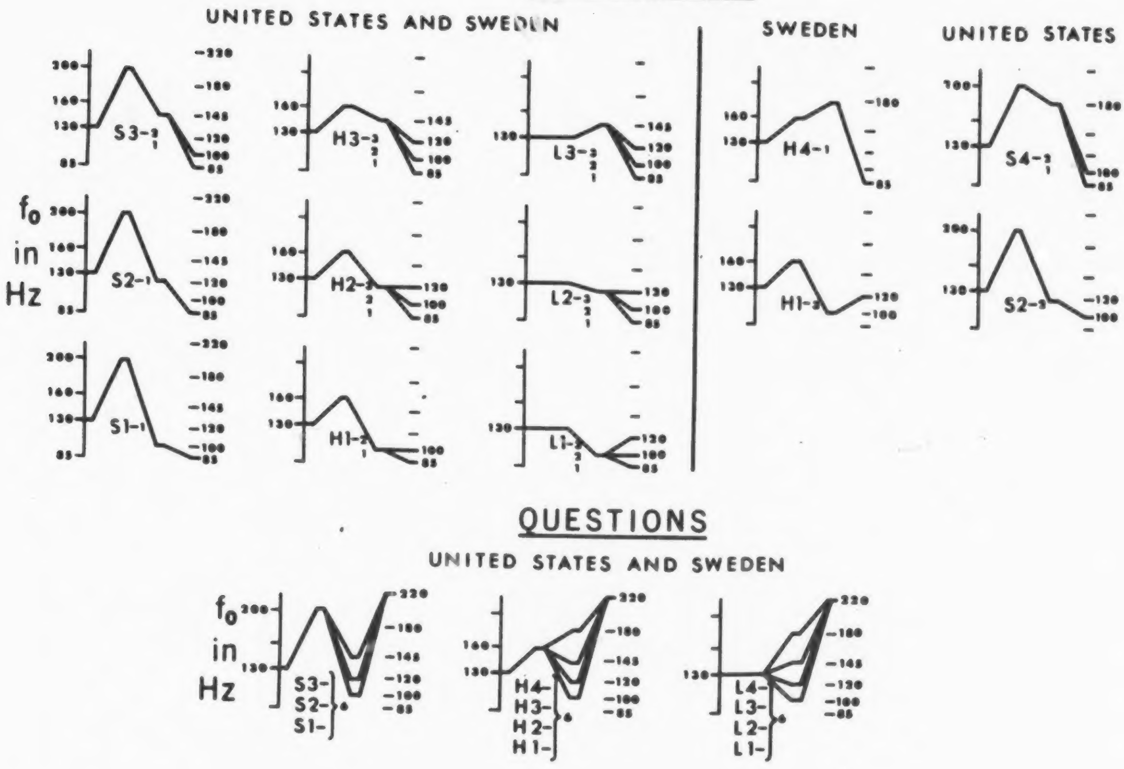
Percentages of Question or Rise Responses (left-axis)
and Statement or Fall Responses (right-axis)
Plotted as Functions of Terminal Glide in Hz



Peak values are constant across rows and turning points are parameters of the curves. For American subjects.

Fig. 5

Schemata of Preferred Statement and Question Contours



Included are all contours for which at least 90% of the judgments of a given language group were in a single category.

Fig. 6

of the utterance. In addition, the contours display at least one of the following: terminal fall, low or middle turning point (1, 2, 3), low or high peak (L, H). The range of preferred contours includes the 2 33 1↓ and 2 22 1↓ contours, suggested as typical by previous observations, but many others are equally acceptable. For example, the superhigh peak, even when followed by a high (S4, US only) or moderately high (S3) turning point, is accepted as a statement provided the terminal fall is large enough; the lower the turning point (i.e., the larger the fall from the peak), the less the needed terminal fall (see S series, Figures 4 and 5). On the other hand, some terminally level contours (H23, H12, L23, L12) and even terminally rising contours (H13, Swedish only; L13) are also accepted as statements. Evidently the terminal fall is not essential, if preceding sections of the contour are low enough (L) or are falling from a moderate level (H).

Broadly, then, peak, turning point, and terminal glide engage in trading relations such that the contour of an acceptable statement has a low to high (rarely, and for US only, superhigh) peak and is, over some portion of its later course, low, falling, or both. (Two anomalous series, H4 and L4, are discussed below under Swedish-U.S. differences.)

Questions. Figure 6 also schematizes contours obtaining 90% or better agreement on a question judgment. For all these contours, the terminal glide is rising and the final pitch of the glide is the highest of the utterance (cf. Uldall, 1962, p. 780; Majewski and Blasdell, 1969). The range of preferred contours includes the expected continuously rising 2 22 $3\uparrow^{4}$ (L36, L46) and 2 33 $3\uparrow^{4}$ (H46) of American English and the superhigh peak contour, 2 44 $2\uparrow^{4}$ (S26) of Swedish, but other contours are also accepted. For example, initially low and falling contours (L1, L2) are heard as questions if the terminal rise is large enough. At the same time, even a terminally level contour (L45, Figures 2-5) gathers more than 80% question judgments from both groups, when the preceding section of the contour has been steadily rising. In fact, this steady rise is a peculiarly powerful question cue that may quite override a large terminal fall that would otherwise cue a statement (cf. H4, L4, discussed below). Again there are trading relations among components of the contour, such that a generally accepted question displays either a rise from peak to turning point (H4, L3, L4) and a relatively small terminal rise, or a fall from peak to turning point and a relatively large terminal rise.

Swedish-U.S. differences. As we have seen, the similarities between Swedish and U.S. judgments are more striking than the differences. The stimulus series included a number of contours presumably unfamiliar to one or other or both groups from their linguistic experience. Yet both groups were able to generalize such contours with more familiar patterns, classifying contours with a relatively high overall pitch as questions, contours with a relatively low overall pitch as statements. Nonetheless, small systematic differences are present.

(1) A comparison of Swedish and U.S. responses to the falling contours of the S2, S3, S4 series (Figures 4 and 5, top left) shows that U.S. subjects tended to give more statement responses than Swedish subjects. The effect is particularly marked for the S4 series on which Swedish statement judgments never reach 90% agreement: a high peak with a high turning point is difficult for Swedish subjects to hear as a statement. This may reflect the fact that Swedish statement intonation shows an earlier fall to a low level after stress

than does English. At the same time, it may be taken as an indirect reflection of a Swedish preference for an overall high contour on questions, so that utterances displaying such a contour are difficult to hear as statements even when completed by a low terminal fall. It is true that the S4 series, which had been expected to collect a large number of question responses due to its overall high level, never obtained 90% agreement on a question judgment from either group. But a control of these items revealed that they gave an impression of protest or indignation rather than of questioning, probably because the low precontour was heard in opposition to the rest of the utterance. A precontour on level 3 might have eliminated this impression and would also have been more similar to what actually occurs in Swedish questions.[7]

(2) As was remarked above, the continuously rising contours (L4 and, to some extent, L3 and H4; see Figures 2 and 3, lower left) were readily accepted by both groups as questions, despite the fact that many of them are unlikely to occur in natural speech. L4, with its low peak rising 50 Hz to the turning point, and H4, with its high peak rising 20 Hz, were preferred to L3 with its low peak rising only 15 Hz. Furthermore, H4 and, especially, L4 elicited relatively few statement responses, even when their terminal glides were falling sharply. U.S. subjects identified these contours as statements even less frequently than the Swedish group. This may reflect the fact that the steadily rising question contour is more widely used in American English than in Swedish and so might be peculiarly difficult for Americans to hear as a statement even when completed by a terminal fall.

In short, the differences between the two groups are small but in directions predictable from linguistic analysis.

Variables Controlling Linguistic Judgments

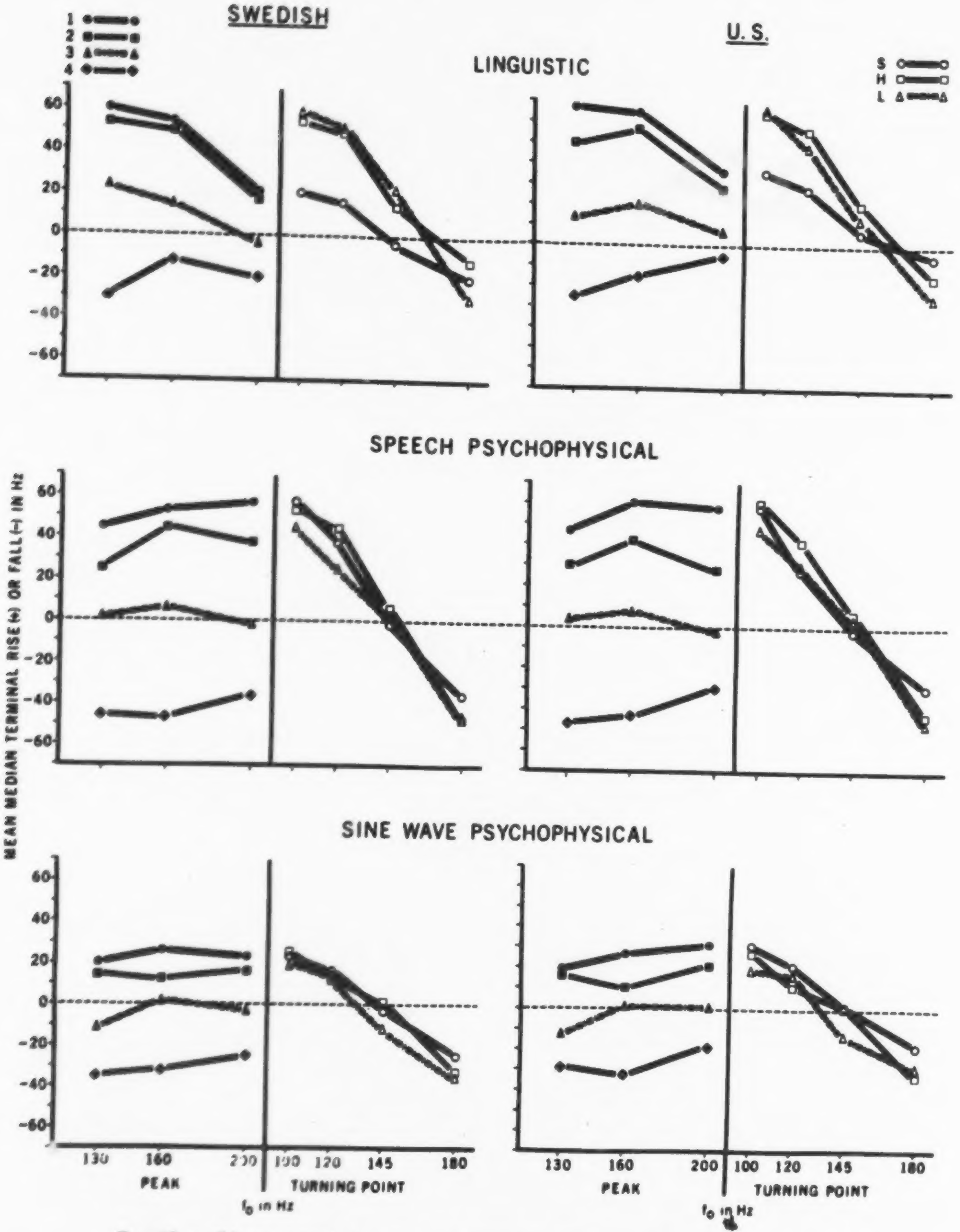
Terminal glide is the single most powerful determinant of linguistic judgments. None of the highly preferred question contours and few of the highly preferred statement contours (Figure 6) lack the appropriate terminal rise or fall. Given a sufficiently extensive terminal glide, earlier sections of the contour have small importance. At the same time, Figures 2-5 show that f_o values at peak and turning point may also play a role.

To provide a consistent criterion for the estimate of peak and turning-point effects, the median of the response distribution for each subject on each series was estimated. The median is the point of subjective equality, the value of the terminal glide at which subjects identify a given contour as a question or a statement 50% of the time. In other words, it is the point of crossover from largely statement to largely question judgments. The means of these medians, or crossover values, for the linguistic judgments are plotted in Figure 7 (row A) for Swedish subjects (left) and U.S. subjects (right).

---

[7] We should probably have included a higher precontour, on level 3, to cover the question contours properly, since the large rise to the highest peak (from level 2 to level 4) gave some contours an unwanted and perhaps dominating effect of protest rather than question (cf. footnote 5.) However, this would have meant a substantial increase in an already lengthy test.

Mean Subject Medians Under the Three Experimental Conditions
for Swedish and American Subjects



In the first and third columns, mean medians are plotted as functions of peak f_o, with turning-point f_o as parameter; in the second and fourth columns, they are plotted as functions of turning-point f_o, with peak f_o as parameter.

Fig. 7

In the first and third plots mean medians are graphed as functions of peak f_o, with turning-point f_o as parameter; in the second and fourth, they are graphed as functions of turning-point f_o, with peak f_o as parameter.

Two cautions should be observed in studying these plots. First, it should be remembered that a median is a single value drawn from the center of its distribution. The relation between the medians of two distributions does not always accurately represent the relations between the upper and lower tails of those distributions. As long as two curves on any plot of Figures 2 to 5 are roughly parallel, the difference between their medians will give a reasonable estimate of their separation along the terminal glide axis. Where there are severe departures from the parallel, the appropriate plots of Figure 7 and of Figures 2 to 5 should be carefully read in conjunction. Second, it should be remembered that the mean of the medians of several distributions is not necessarily equal to the median of the combined distribution. Since the values of Figure 7 are the means of subject medians, they do not always agree exactly with the group median values read from Figures 2 to 5.

With these precautions in mind we return to row A of Figure 7. If the direction of the terminal glide were the sole determinant of linguistic judgments, we would expect all crossover values to fall at zero, the level of the dashed horizontal lines across Figure 7. In fact, crossover values deviate considerably from zero: both the direction and the extent of their deviation vary with peak and turning point.

The peak effect (plots 1 and 3) is the smaller. For neither Swedish nor U.S. subjects does a change of peak f_o from 130 Hz to 160 Hz (from L to H) have any consistent, significant effect. But a change from 160 Hz to 200 Hz (from H to S) does reliably reduce the crossover value for all contours, except that having a turning point at 180 Hz for the U.S. group. (This reversal is probably not reliable, as study of the bottom left plot of Figure 3 will suggest.) These effects are statistically significant by matched pair t-tests between medians for turning points 1, 2, and 3 in both groups ($p<.05$). They may be clearly seen in the left columns of Figures 2 and 3. Reading down the columns we note the leftward separation of the S curves. The separation is reduced for turning point 3 and gives place to the L curve, with its steadily rising contour, for turning point 4. We may also note that, as the terminal rise increases, the peak effect in the upper three plots disappears. In short, if the turning point is at a low to middle f_o and the terminal rise is slight, a very high (level 4) peak at the stress leads to a significant increase in the number of questions heard and, by corollary, to a significant decrease in the number of statements.

The turning-point effect (plots 2 and 4 of Figure 7) is both larger and more consistent than the peak effect. For all values of peak f_o, an increase in turning-point f_o is associated with a decrease in crossover value. The decrease is significant by matched pair t-tests between medians ($p<.05$) for all turning-point shifts, except those from 100 to 120 Hz for the Swedish S, H, and L curves and for the U.S. S and H curves. The effect is also considerably reduced, if the contour has a peak at 200 Hz (S). (See top left plots of Figures 4 and 5.) This again suggests that the high peak alone is a powerful question cue for both language groups.

Psychophysical Judgments

Speech Waves

Psychophysical judgments of the speech-wave terminal glides differ from and resemble linguistic judgments of the entire utterance in important ways. The main difference may be seen in the center columns of Figures 2 and 3: the effect of the high peak is absent from the Swedish data and much reduced in the U.S. data. The main similarity may be seen in the center columns of Figures 4 and 5: the turning-point effect is present and even more pronounced than in the linguistic judgments.

Figure 7 (row B) summarizes the data. The peak effects (plots 1 and 3) are inconsistent. An increase in peak f_0 from 130 Hz (L) to 160 Hz (H) yields in every instance, except the high turning-point series for Swedish subjects, an increase rather than a decrease in the crossover value of the terminal rise. Two of these increases (for turning points 1 and 2) are significant for both groups ($p<.05$ by a matched pair t-test between medians). On the other hand, an increase of peak f_0 from 160 Hz to 200 Hz yields, for the Swedish subjects, two increases and two decreases, none of them significant. The absence of a consistent peak effect for the Swedish subjects is evident in the middle column of Figure 2. For the U.S. subjects, the picture is somewhat different: crossover values decrease from H to S for turning points 1, 2, and 3 and increase for turning point 4, exactly as in the linguistic data. The effects are reduced and statistically significant only for turning point 2. But a trend is present and quite evident in the middle column of Figure 3.

The turning-point effect, on the other hand (center columns of Figures 4 and 5; plots 2 and 4, row B of Figure 7) is similar to and even more pronounced than the corresponding effect in the semantic data. All shifts are significant by matched pair t-tests ($p<.05$), except that from turning point 1 to 2 in the Swedish L series. For both groups, the higher the turning point, the smaller the terminal rise needed for a rise to be consistently heard. The similarity to the linguistic results is most marked for the H and L series (second and third rows, Figures 4 and 5): H4 and L4 are again anomalous series, readily heard as rising even when the terminal glide is falling. In the S series the turning-point effect is even more pronounced than for the linguistic judgments.

Sine Waves

From the steepened functions of Figures 2 to 5 (right-hand columns) it is evident that subjects were in better agreement on their sine-wave than on their speech psychophysical or linguistic judgments. The two language groups are also in close agreement, which gives some confidence that the differences between their linguistic judgments are reliable.

Figures 2 and 3 (right-hand columns) show that the effect of the high peak is absent. As in the speech psychophysical data, low peak contours tend to be the most accurately judged, particularly by the Swedish. But the effects are neither fully consistent nor statistically significant (see plots 1 and 3, row C, Figure 7).

On the other hand, the turning-point effects (plots 2 and 4, row C, Figure 7) are clear, similar to those observed in the linguistic and speech psychophysical data but considerably reduced. The effects are significant by matched pair t-tests ($p<.05$) for all turning-point shifts, except those from 100 Hz to 120 Hz for the S and H curves in both groups, and may be seen in the right-hand columns of Figures 4 and 5. Note that H4 and L4 are no longer anomalous series.

DISCUSSION

Cross-language comparisons. There are striking similarities between Swedish and U.S. judgments of these intonation contours. Despite small, linguistically predictable differences, both groups tend to classify contours with a high peak or terminal rise as questions, contours with a low peak or terminal fall as statements. Hermann (1942) has pointed out the generality across languages, including Swedish, of a high pitch for questions (see also Hadding-Koch, 1961, especially pp. 119 ff.). Bolinger (1964), among others, has discussed the apparently "universal tendency" to use a raised tone to indicate points of "interest" within utterances and also to indicate that more is to follow, as in questions (cf. Hadding-Koch, 1965). The data of this experiment are consistent with these "universal tendencies."

Perceptual relations within a contour. We are now in a position to resolve some of the uncertainties left by our previous study. Consider, first, the turning-point effect. Since this is present and significant under all three experimental conditions, we must assign it auditory status and assume that it takes linguistic effect indirectly by altering subjects' perceptions of the terminal glide. Furthermore, since it is present, even though reduced, in the sine-wave data, our account of the process by which it affects perception of the terminal glide cannot invoke specialized mechanisms peculiar to speech.

We may gather some idea of the process from a study of plots 2 and 4 in row B, Figure 7 or of the center plots in Figures 4 and 5. The terminal glide of a contour, such as H1, with a strong fall from peak to turning point (160 Hz to 100 Hz) requires a terminal rise of about 50 Hz if it is to be judged 50% of the time as rising; while the terminal glide of a contour, such as H4, with a steady rise for more than 200 msec before the terminal glide, is heard as rising 50% of the time, even when the glide is falling by about 50 Hz. Evidently listeners have difficulty in separating the terminal glide from earlier sections of the contour, if those earlier sections have a marked movement. The terminal glides of contours with a turning point (145 Hz in S3, H3, L3) close to the precontour level of 130 Hz are more accurately perceived: the median values are close to zero in every plot of Figure 7, columns 2 and 4. Listeners are perhaps able to average across earlier sections of such contours and establish an anchor against which terminal glide may be judged.

All this implies that later sections of the contours in this study (that is, roughly the last 400 msec, from peak to turning point to end point) were processed by listeners as a single unit, with attention focussed on the terminal glide. If a listener was able to separate the glide perceptually from the immediately preceding section (as in the S3, H3, L3 series), his linguistic judgments followed pretty well the traditional formulation of rise for questions, fall for statements. If he was not able to separate the glide, due to the difficulty--heightened perhaps for a complex speech signal--of tracking a rapidly modulated frequency, relatively gross movements of the terminal glide were necessary for him to be sure whether he had heard a rise or a fall, a question or a statement.

Interpretation of the peak effect is more difficult. In our earlier study, the effect was clear in both linguistic and psychophysical judgments of both groups, though the Swedish were less consistent in their psychophysical judgments than the Americans. In this study, a peak effect is significantly present in linguistic judgments, totally absent from sine-wave judgments, and for speech psychophysical judgments, marginally present only for the Americans.

We will consider the speech psychophysical data below. Here, the important point is that the peak effect is reliably present in the linguistic, but absent from the sine-wave, judgments. We may therefore, with reasonable certainty, reject an auditory (or psychophysical) account and assign a direct linguistic function to the peak. Unlike turning-point variations, peak variations do not take linguistic effect by altering listeners' perceptions of the terminal glide. Rather, the peak is a distinct element to be weighed with the perceived terminal glide in determining the linguistic outcome.

We should note, in caution, that peak and terminal glide are not always simply additive in their effects. For example, a contour with a steady rise from precontour to end point may require a relatively small terminal rise to be heard as a question, despite its low peak (e.g., L3 series). Here, it seems to be the overall sweep of the pattern that determines the judgment rather than the frequency levels of particular segments of the contour.

However, with few exceptions, two factors would seem to govern linguistic judgments of intonation contours, such as those of this study: fundamental frequency at the peak and perceived terminal glide. The entire contour is then interpreted as a unit with these factors in weighted combination, and with the heavier weight being assigned to the terminal glide. If a terminal fall is heard, the listener interprets the utterance as a statement, unless the fall was slight and he has also heard a very high peak; if a terminal rise is heard, the listener interprets the utterance as a question, unless the rise was slight and he has also heard an unusually low peak (cf. Greenberg, 1969, Ch. 2; Ohala, 1970, pp. 101 ff.).

Auditory-linguistic interactions. We turn, finally, to the speech psychophysical data. Our problem is to understand the instances in which speech psychophysical judgments follow the linguistic more closely than the sine-wave judgments. Obviously, these instances can only occur where linguistic judgments of the entire contour differ from auditory judgments of the terminal sine-wave glide, that is, where the contour carries some linguistically relevant cue other than terminal glide. For questions, such cues include a superhigh peak or a monotonic rise from precontour to turning point. Accordingly we find a tendency for speech psychophysical judgments to follow linguistic judgments in the superhigh (S) peak series (see Figure 3) and in the high turning-point series (see Figures 4 and 5). Consider, particularly, the results for speech contours of the H4 and L4 series. Listeners in both groups often judge these contours both as questions and as terminally rising, even though they are able to hear that the corresponding sine-wave contours have terminal falls. Since listeners cannot have judged the contours to be questions

cause they heard a terminal rise, we are tempted to conclude that they
ard the terminal rise because they judged the contours to be questions:
nguistic decision determined auditory shape.

Before elaborating on this, it is important to remark that such effects
not always occur where they might be expected. For example, the peak ef-
ct was clearly present in the speech psychophysical judgments of both groups
our earlier study but is reduced to a marginal effect in the American and
s disappeared entirely from the Swedish speech psychophysical data of the
esent study. We can hardly therefore call on the effect to support a gen-
al account in terms of some specialized perceptual mechanism, such as that
oposed by Lieberman (1967). At the same time, the results are evidently
culiar to speech and cannot be handled in purely auditory terms. What we
ed, therefore, is an account in terms of a process that may vary with ex-
rimental conditions and subjects.

An interesting hypothesis, suggested above, is that the results reflect
e blend of serial and parallel processing that characterizes the perception
spoken language (and of other complex cognitive objects) (cf. Fry. 1956;
istovich et al., 1968; Studdert-Kennedy, in press). We may conceive the
rceptual process as divided into stages (auditory, phonetic, phonological,
c.), but we must also suppose there to be feedback from higher to lower
vels which may serve to correct or verify earlier decisions. Perceptual
orrection" of an auditory or phonetic decision, in light of a higher lin-
istic decision, will presumably not occur if the lower decision is firm.
herwise, we would not be able to deem the intonation of an actor "wrong"
to understand a speaker, yet perceive his dialect to be unfamiliar. How-
r, in difficult listening conditions and under certain, as yet undefined,
oustic conditions, perceptual "correction," sufficient to produce a com-
lling phonetic illusion, may occur (Miller, 1956). Warren (1970; Warren
Obusek, 1971) has shown that listeners may clearly perceive a phonetic
ment that has been excised from a recorded utterance and replaced by an
raneous sound (cough, buzz, tone) of the same duration. The important
nt is that listeners perceive the correct segment: the precise form of the
netic illusion is determined not by the acoustic conditions alone but also
higher-order linguistic constraints.

Here, the illusion is auditory rather than phonetic, but a similar me-
nism may be at work. Asked to interrupt his normal perceptual process at
rephonetic auditory stage, the listener falls back on his knowledge of the
guage. As we have seen, the single most powerful cue for question/state-
t judgments in this experiment was the terminal glide. Listeners evident-
prefer, and presumably expect, a question to end with a rise, a statement
h a fall (see Figure 6). However, earlier sections of the contour may also
er into the decision and, if sufficiently marked, override an incompatible,
relatively weak, terminal glide. Called upon to judge this glide, the
tener then assigns it a value consonant with his linguistic decision. That
to say, if other factors dominate his linguistic decision, he may be led
o nonveridical perception of the terminal glide.

The degree to which this happens might be expected to vary with the re-
ive strengths of the cues controlling linguistic decision. And in fact,
as the peak effect in the linguistic data was stronger for our first
dy than for our second, so too was the peak effect in the speech psycho-
sical data. Similarly, just as the question cue in the rising contours of

the H4 and L4 series is stronger for the Americans than for the Swedish, so too is the tendency toward nonveridical judgment of the terminal glide.

However, we should not expect to be able to develop a fully coherent account of our results in these terms, since we are ignorant of the limiting linguistic and acoustic conditions of the illusion. We are currently planning to broaden our understanding of the effect by taking advantage of what is known about the various acoustic cues to word stress (Fry, 1955, 1958). We might expect, for example, that, if linguistic decision can indeed determine auditory shape, syllables of equal duration, judged to be differently stressed on the basis of differences in either intensity or fundamental frequency, would also be judged of unequal length. The ultimate interest of the account is in its suggestion that the auditory level is not independent of higher levels but is an integral part of the process by which we construct our perceptions of spoken language.

REFERENCES

Allen, G.A. (1971) Acoustic level and vocal effort as cues for the loudness of speech. J. acoust. Soc. Am. 49, 1831-1841.

Bolinger, D.L. (1958) A theory of pitch accent in English. Word 14, 109-149.

Bolinger, D.L. (1964) Intonation as a universal. *Proc. IXth Intl. Cong. Linguistics, Cambridge, Mass., 1962*. (The Hague: Mouton and Co.) pp. 833-848.

Chistovich, L.A., Golusina, A., Lublinskaja, F., Malinnikova, T. and Zukova, M. Psychological methods in speech perception research. Z. Phon. Sprachwiss. u. Komm. Fschg. 21, 102-106.

Chomsky, N. and Halle, M. (1968) *The Sound Pattern of English* (New York: Harper and Row).

Cooper, F.S. (1965) Instrumental methods for research in phonetics. *Proc. Vth Intl. Cong. Phonetic Sci., Munster, 1964*. (Basel, S. Karger) pp. 142-171.

Delattre, P. (1963) Comparing the prosodic features of English, German, Spanish and French. IRAL 1:3-4, 193-210.

Denes, P. (1959) A preliminary investigation of certain aspects of intonation. Language and Speech 2, 106-122.

Fourcin, A.J. (in press) Perceptual mechanisms at the first level of processing. In *Proc. VIIth Intl. Cong. Phon. Sci., Montreal, 1971*.

Fry, D.B. (1955) Duration and intensity as physical correlates of linguistic stress. J. acoust. Soc. Am. 27, 765-768.

Fry, D.B. (1956) Perception and recognition in speech. In *For Roman Jakobson* M. Halle, H.G. Lunt, H. McClean, and C.H. van Schooneveld, eds. (The Hague: Mouton) pp. 169-173.

Fry, D.B. (1958) Experiments in the perception of stress. Language and Speech 1, 126-152.

Fry, D.B. (1968) Prosodic phenomena. In *Manual of Phonetics*, B. Malmberg, ed. (Amsterdam: North Holland Publ. Co.) pp. 365-410.

Gårding, E. and Abramson, A.S. (1965) A study of the perception of some American English intonation contours. Studia linguistica XIX, 61-79.

Greenberg, S.R. (1969) An experimental study of certain intonation contrasts in American English. UCLA Working Papers in Phonetics 13.

Hadding-Koch, K. (1961) *Acoustico-Phonetic Studies in the Intonation of Southern Swedish*. (Lund: Gleerups).

Hadding-Koch, K. (1965) On the physiological background of intonation. Studia linguistica XIX, 55-60.

Hadding-Koch, K. and Studdert-Kennedy, M. (1963) A study of semantic and psychophysical test responses to controlled variations in fundamental frequency. Studia linguistica XVII:2, 65-76.

Hadding-Koch, K. and Studdert-Kennedy, M. (1964) An experimental study of some intonation contours. Phonetica 11, 175-185.

Hadding-Koch, K. and Studdert-Kennedy, M. (1965) Intonation contours evaluated by American and Swedish test subjects. Proc. Vth Intl. Cong. Phonetic Sci., Munster, 1964. (Basel: S. Karger) pp. 326-331.

Hermann, E. (1942) Probleme der Frage. Nachrichten von der Akademie der Wissenschaften in Gottingen 3-4.

Hockett, C.F. (1955) A manual of phonology. Indiana Univ. Publ. in Anthropology and Linguistics 11.

Ladefoged, P. and McKinney, N.P. (1963) Loudness, sound pressure and subglottal pressure in speech. J. acoust. Soc. Am. 35, 454-460.

Lehiste, I. (1970) Suprasegmentals. (Cambridge, Mass.: The M.I.T. Press).

Léon, P.R. (in press) Où en sont les recherches sur l 'intonation. In Proc. VIIth Intl. Congr. Phon. Sci., Montreal, 1971.

Lieberman, P. (1967) Intonation, Perception, and Language. (Cambridge, Mass.: The M.I.T. Press).

Lieberman, P. (in press) A study of prosodic features. In Current Trends in Linguistics, Vol. XII, T.A. Sebeok, ed. (The Hague: Mouton). (Also in Haskins Laboratories SR-23, 1970).

Majewski, W. and Blasdell, R. (1969) Influence of fundamental frequency cues on the perception of some synthetic intonation contours. J. acoust. Soc. Am. 45, 450-457.

Miller, G.A. (1956) The perception of speech. In For Roman Jakobson, M. Halle, H.G. Lunt, H. McClean and C.H. van Schooneveld, eds. (The Hague: Mouton) pp. 353-359.

Ohala, J. (1970) Aspects of the control and production of speech. UCLA Working Papers in Phonetics 15.

Pike, K.L. (1945) The Intonation of American English. (Ann Arbor: Univ. of Michigan Press).

Studdert-Kennedy, M. (in press) The perception of speech. In Current Trends in Linguistics, Vol. XII, T.A. Sebeok, ed. (The Hague: Mouton). (Also in Haskins Laboratories, Status Report on Speech Research SR-23, 1970.)

Uldall, E.T. (1962) Ambiguity: Question or statement? or "Are you asking me or telling me?" In Proc. IVth Intl. Cong. Phonetic Sci., Helsinki, 1969 (The Hague: Mouton) pp. 779-783.

Warren, R.M. (1970) Perceptual restoration of missing speech sounds. Science 167, 392-393.

Warren, R.M. and C.J. Obusek. (1971) Speech perception and phonemic restorations. Perception and Psychophysics 9 (3B), 358-362.

Glottal Modes in Consonant Distinctions[*]

Leigh Lisker[+] and Arthur S. Abramson[++]
Haskins Laboratories, New Haven

Our most direct knowledge of how the larynx operates derives from observations by means of a laryngeal mirror inserted through the open mouth, from which we know that voicing involves adduction of the arytenoids so that the vibrating vocal folds are closely, but not tightly, approximated, that quiet respiration is accomplished with the glottis well opened, and that whisper, creaky voice, falsetto, murmur, "glottal stop," and "aitch" involve still other more or less easily distinguished modes of laryngeal adjustment. The observational method is, of course, not applicable to speech, and up till fairly recently whatever was said about the functioning of the larynx during speech was by inference, and subject in part to controversy. It was supposed, very plausibly, that during voiced intervals in which the mouth is open the larynx operates just as observed during the phonation of prolonged vowel-like sounds. There was less agreement, and sometimes less certainty, as to laryngeal functioning during voiceless intervals in running speech, which typically coincide more or less with constriction of the supraglottal airway. Given the structure of the vocal tract and the myoelastic-aerodynamic theory of phonation, and assuming the larynx fixed in the voicing mode, we should expect a more or less apid extinction of voicing to be inevitable when there is severe constriction. Conversely we should expect the suppression of voicing only in that circumstance. Compatible with this is the observation sometimes made that sounds with little constriction are "normally" voiced, and its less often stated corollary that bstruents, particularly stops, are "normally" voiceless. If a language is normal" in this way, then it seems reasonable to suppose that in fact a single lottal mode, that of voicing, is maintained without significant change in tterances of that language, with shifts in mode reserved for paralinguistic ffects. The absence of a distinctive voicing feature is then matched by the bsence of differential control of the larynx during speech. But while such anguages are reported, they are not very common. The literature of phonetic escription suggests, rather, a special affinity between voicing as a distinc-ive feature and stop consonants, so that voiced stops are by no means rare. f we suppose that the voicelessness of certain stops is compatible with the lottal mode appropriate to voicing, then the presence of voicing in others mplies some other mode and/or some other way of maintaining the necessary ransglottal airflow during occlusion. Theoretical arguments have been advanced Halle and Stevens, 1967) for a shift in glottal mode as a necessary condition or stop voicing, a shift which effects a reduction in the resistance to air- low through the glottis. Moreover, if voicelessness persists after the stop elease, as in the case of voiceless aspirates, then still another mode of lottal adjustment would be implicated.

---

[*] Paper presented at the Seventh International Congress of Phonetic Sciences, Montreal, 1971. To be published in the Proceedings.

[+] Also University of Pennsylvania, Philadelphia.

[++] Also University of Connecticut, Storrs.

It has been further asserted that, in addition to mode of glottal adjust- ent, a dimension of articulatory force plays a strong role in determining hether or not vocal-fold vibration accompanies a supraglottal constriction Chomsky and Halle, 1968). This fortis-lenis dimension has been variously nderstood; currently it is the fashion to say that it determines the extent o which the pharynx is free to expand in response to an increase in air ressure such as occurs during obstruent production. Obviously a transglottal irflow can be better maintained during an occlusion if the pharyngeal cavity olume is increasing, and Rothenberg (1968) has reported experiments measuring he compliance of the cavity walls which yield values compatible with the dura- ions of voiced closure observed in speech. In the case of aspiration, more- ver, still another parameter, subglottal air pressure, has been enlisted by omsky and Halle (1968) as a significant factor by way of explaining the latively high rates of airflow observed.

By and large, much of what is said to be known about the management of op voicing and aspiration is more hypothetical than data-based, and where ere are data, they are more often than not derived from nonsense exercises the speech mechanism whose relation to running speech is not clear. With cent developments in instrumentation, new techniques have come into use ich yield more direct information on the glottis in consonant production. udies in transillumination, electroglottography, electromyography, and fiber- tics and X-ray cinephotography have already provided some findings that fail confirm some of the recently stated theories of glottal behavior as it relates distinctive voicing. From transillumination and fiberoptics studies carried at Haskins Laboratories, for example, it appears that voiceless unaspirated ops, in English at least, most often involve some opening movement of the ytenoids, while on the other hand there is no detectible movement of these rtilages in a large majority of voiced stops observed (Sawashima et al., 70). If a shift in glottal mode is in theory required for stop voicing, d if it is superfluous for the voiceless unaspirated stops, then it is zzling that evidence of a special glottal adjustment in the first case is so usive and in the second seems so clear. If there is, in fact, a gesture of voicing rather than to ensure voiced occlusion, it might be inferred that a rtis-lenis difference is of less than crucial importance, at any rate for uent American English. Nor has there been any demonstration that higher sub- ottal pressures are required for aspiration, while there is clear evidence t the area of glottal aperture at the time of stop release is directly re- ted to the prominence of this feature. The mechanism by which aspiration, something much akin to it, is produced during the release of voiced stops not well studied. It seems possible, though, that this variety of aspira- n is voiced, unlike the more commonly found aspiration, simply because the ttal aperture does not become large enough for vocal-fold vibration to cease the absence of an articulatory constriction.

In summary, it seems to us that theories of stop voicing and aspiration t stress the importance of extralaryngeal factors can claim less basis in erved fact than does one which stresses the paramount role of the larynx, cifically the positioning of the arytenoid cartilages as it determines

ottal aperture. It is difficult to deny that extralaryngeal factors may
fect voicing singnificantly, but it is one thing to argue that they have
e capability, another to demonstrate that they do in fact regularly operate
a manner consistent with that capability. Glottal adjustment alone does
t determine the voicing state of a stop consonant, but no other factor
ems to be nearly as important.

REFERENCES

omsky, N. and M. Halle (1968) The Sound Pattern of English. (New York: Harper and Row).

lle, M. and K.N. Stevens (1967) On the mechanism of glottal vibration for vowels and consonants. Quarterly Progress Report of the Research Laboratory of Electronics, M.I.T., 85, 267-271.

, C-W. (1970) A theory of aspiration. Phonetica 21, 107-116.

henberg, M. (1968) The breath-stream dynamics of simple-released-plosive production. Bibliotheca Phonetica 6, 92-94 (Basel: S. Karger).

ashima, M., A.S. Abramson, F.S. Cooper, and L. Lisker (1970) Observing laryngeal adjustments during running speech by use of a fiberoptics system. Phonetica 22, 193-201.

Voice Timing in Korean Stops[*]

Arthur S. Abramson[+] and Leigh Lisker[++]
Haskins Laboratories, New Haven

Linguists have disagreed over the features distinguishing the three manner categories of Korean plosives. The three categories of labial, apical, and dorsal stops and palatal affricates are variously described for initial position, using one or more of the following terms: I. Voiceless, tense, long, and glottalized; II. Voiceless, lax, and slightly aspirated; III. Voiceless, heavily aspirated, and lax by some but tense by others. A further complication is the frequent voicing of Category II in a medial voiced environment.

We have devoted much of our research effort to questions of laryngeal control in stop consonants. We have shown that various conditions of voicing and aspiration in word-initial stops in a wide variety of languages depend upon differences in voice-onset time (VOT), the temporal relation between stop release and onset of glottal pulsing (Lisker and Abramson, 1964). Some aspects of the conflicting descriptions of Korean plosives suggested that we test the efficacy of VOT in that language. Combining data from our 1964 study with some recent additions, we present VOT measurements for two native Korean speakers' initial apical stops in Figure 1. The abscissa shows VOT in intervals of 10 msec; zero represents the moment of stop release. The ordinate shows the frequency distribution of VOT values for each of the three categories. Although Speaker B tends to have slightly higher values, the overall results are quite comparable for the two speakers and for the labial and dorsal stops not shown in the figure. Category III is well separated from the others, but I and II overlap somewhat. Similar data have been published by others (Kim, 1965; Han and Weitzman, 1970). Of course, where II assimilates to preceding voicing in medial position, VOT separates all three categories.

The foregoing mixed results made us wonder to what extent VOT might provide sufficient perceptual cues for discriminating the three categories. Also, having shown the perceptual efficacy of VOT for Thai, Spanish, and English (Lisker and Abramson, 1970), we wished to extend our comparative phonetic investigation of the dimension to Korean by studying perception as well as production. Lest we later find instability in the phonological distinctions of concern to us, we proved that randomized words differing only in initial stop categories could be identified with ease. We then exposed native speakers to a continuum of synthetic VOT variants ranging from a voicing lead of 150 msec before the release of the stop to a voicing lag of 150 msec after the release, for identification as

---

[*] Paper presented at the Seventh International Congress of Phonetic Sciences, Montreal, 1971. To be published in the Proceedings.

[+] Also University of Connecticut, Storrs.

[++] Also University of Pennsylvania, Philadelphia.

KOREAN INITIAL STOPS
(Apical)

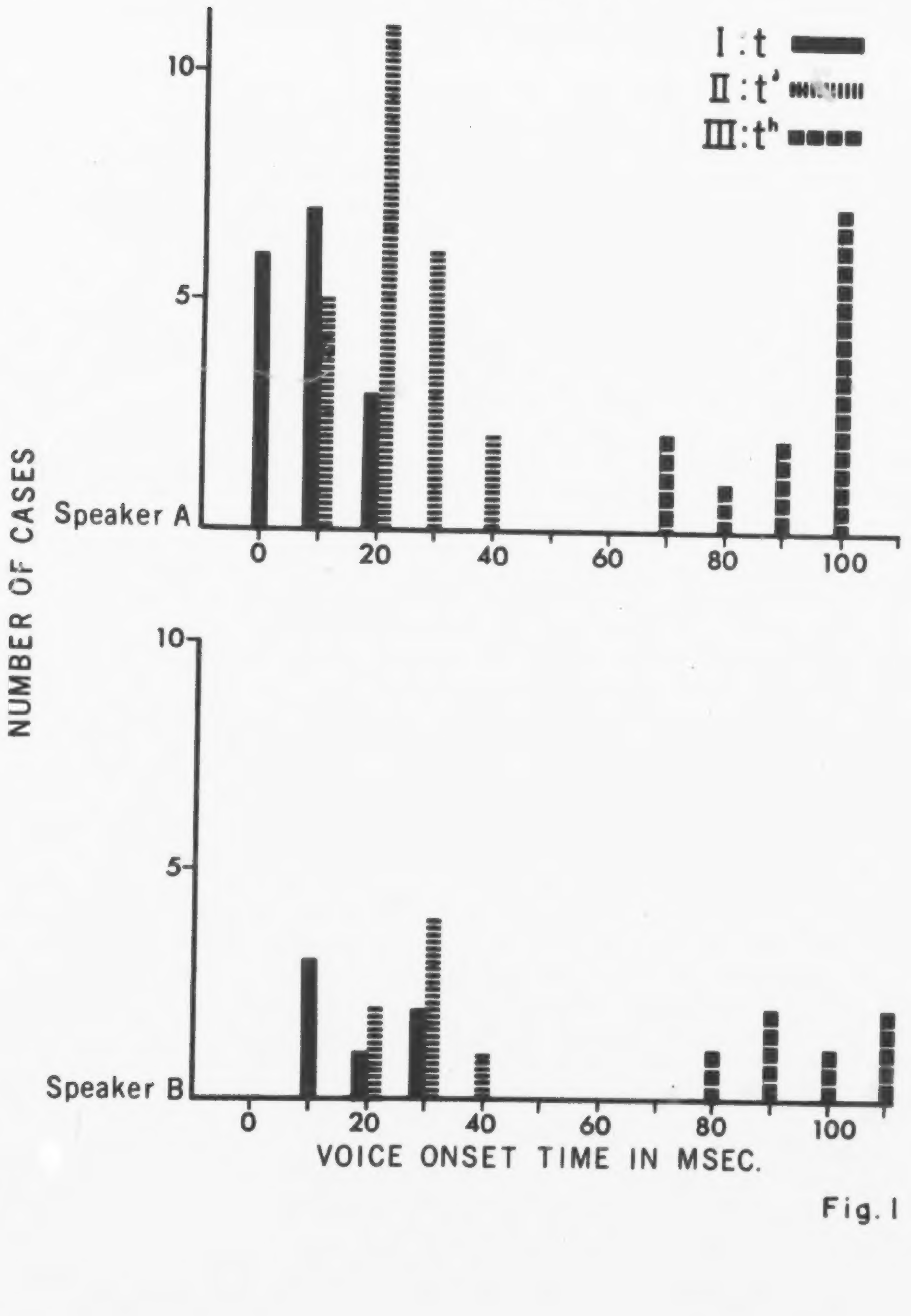


Fig. I

Korean syllables at each place of articulation. There were two experimental conditions: (1) a restricted range with all voicing lead variants excluded, thus apparently simulating spoken Korean; (2) the full continuum, thus including variants found only in non-initial position in the spoken language. The range is divided into 10-msec steps except for the portion from a lead of 10 msec to a lag of 50 msec, which is divided into 5-msec steps.

We present labelling responses for the synthetic apical stops only, but the data are typical of all three places of articulation. Figure 2 contains the identification curves for the restricted range. Values of voicing lag are indicated along the abscissa and percent identification as each stop along the ordinate. The left half of the figure is blocked out to show that no lead variants were used. The five subjects responded to the stimuli in three ways. At the top of the figure we see that HL called the variants from 0 to 50 msec Category I and the rest, Category III; he heard none as II. The middle display shows a partition of the range into I, II, and III, in that order; these three subjects, then, behaved much as if VOT were a straightforward cue. At the bottom of the figure, YH divides most of the stimuli between II and III, while weakly favoring I only at 60 and 70 msec.

The responses to the full continuum, including the lead variants marked with negative VOT values, are given in Figure 3. Three response patterns are shown by the four subjects. At the top of the figure, BC simply divides the range into I, II, and III, but with occasional labelling of lead variants as II. By and large, she would seem to hearing voicing lead as a badly pronounced version of the unaspirated stop. We can perhaps understand her vacillation by looking at the middle of Figure 3. There we see two subjects who yielded the startling result that only variants with voicing lead were heard as II, while the rest of the continuum was divided between I and III. It should be recalled that audible laryngeal pulsing does not occur during initial stop occlusions in Korean; therefore, the obvious interpretation of our data is that, upon detecting such abnormal voicing, at least some Koreans feel they must assign it to the one category that has voiced occlusions in any context at all. This implies that they are somehow aware of glottal pulsing, or the underlying laryngeal gesture, as a component of II. At the bottom of Figure 3, CH not only does the same thing but also assigns several slightly aspirated variants--those from 35 to 80 msec of lag--to Category II.

The complicated response patterns and production data lead us to two inferences: (1) the timing of glottal adjustments relative to supraglottal articulation contributes to the Korean distinctions, and (2) there must be another dimension that works with VOT in distinguishing the categories. An accumulation of acoustic data on the matter has been furnished by Han and Weitzman (1970), and Kim (1965) in addition to Kim's (1965, 1970) physiological data. We are tempted to believe that the difficult question of the distinction between Categories I and II in initial position will be resolved by further examination of laryngeal mechanisms. Recent fiberoptics work by Kagaya (1971) supports this belief. Also, some speakers have quite audible vocal fry or laryngealization in Category I. We plan to take a close look at this phenomenon by means of our fiberoptics system.

KOREAN LABELLING JUDGMENTS
(Stimulus Range: 0/+150 msec.)

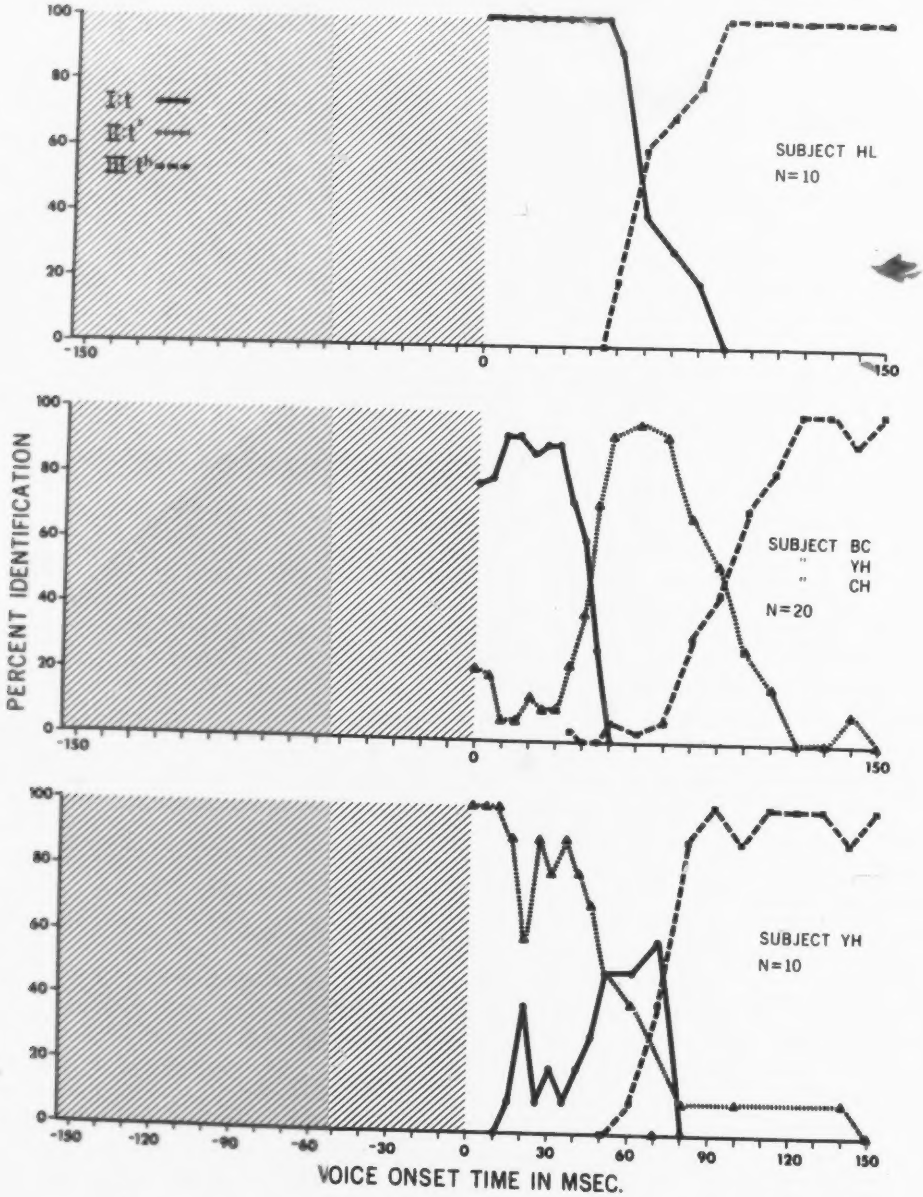


Fig. 2

KOREAN LABELLING JUDGMENTS
(Stimulus Range: -150/+150 msec.)

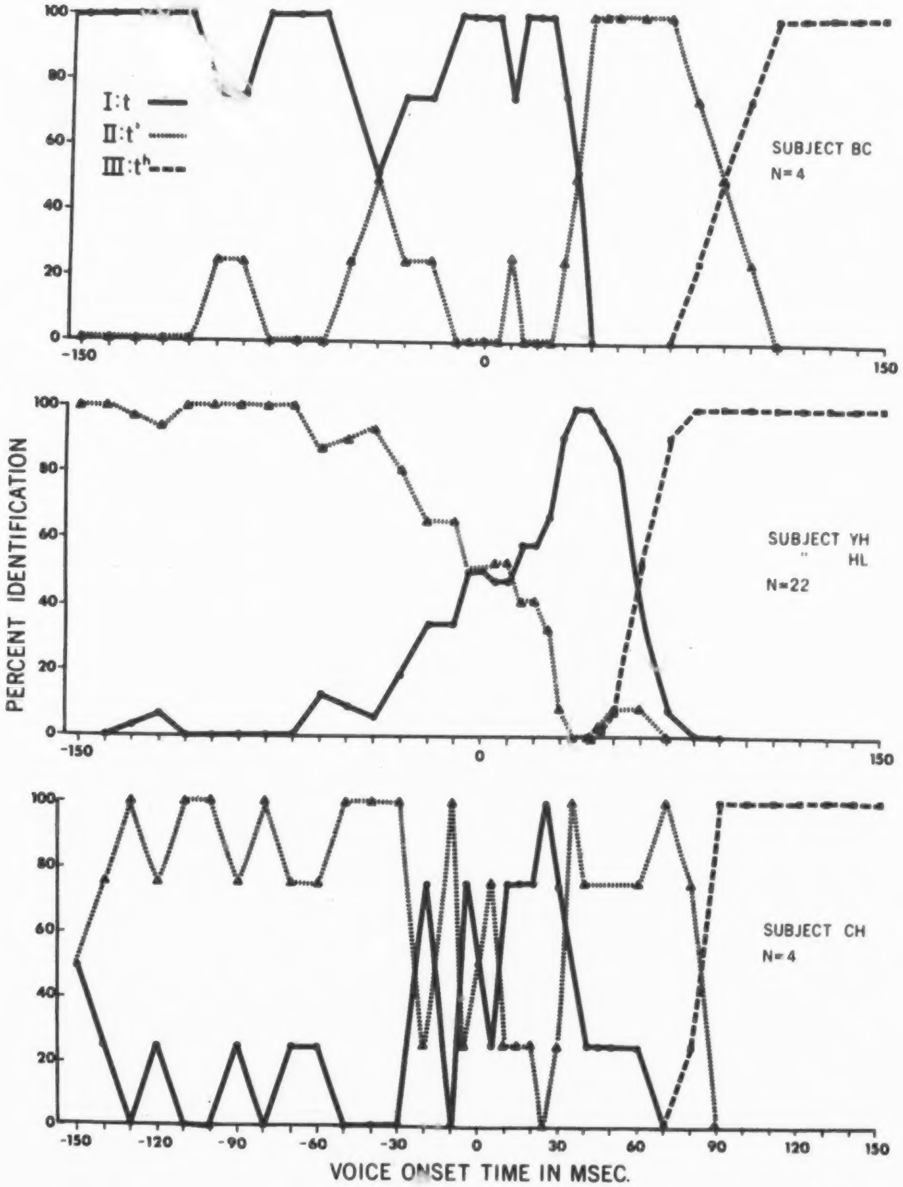


Fig. 3

REFERENCES

Han, M.S. and R.S. Weitzman (1970) Acoustic features of Korean /P,T,K/, /p,t,k/ and /p^h,t^h,k^h/. Phonetica *22*, 112-128.

Kagaya, R. (1971) Laryngeal gestures in Korean stop consonants. Res. Inst. of Logopedics and Phoniatrics, U. of Tokyo, Annual Bull. *5*, 15-23.

Kim, C-W. (1965) On the autonomy of the tensity feature in stop classification (with special reference to Korean stops). Word *21*, 339-359.

Kim, C-W. (1970) A theory of aspiration. Phonetica *21*, 107-116.

Lisker, L. and A.S. Abramson (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word *20*, 384-422.

Lisker, L. and A.S. Abramson (1970) The voicing dimension: Some experiments in comparative phonetics. In *Proc. 6th Intl. Cong. Phon. Sci., Prague 1967*, 563-567 (Prague: Academia).

Interactions Between Linguistic and Nonlinguistic Processing[*]

Ruth S. Day[+] and Charles C. Wood[++]

ABSTRACT

Possible interactions between the processing of linguistic and nonlinguistic stimulus dimensions were examined by selecting combinations of the following stimuli for use in two-choice identification tasks: /ba/ at a low fundamental frequency, /ba/-high, /da/-low, /da/-high. Linguistic Task--Subjects identified which stop consonant was present on each trial. In the One-Dimension Condition the stimuli were /ba/-low and /da/-low; hence only stop consonants varied. In the Two-Dimension Condition all four stimuli were presented; again subjects identified which stop consonant occurred but had to ignore variation in the irrelevant dimension, fundamental frequency. Nonlinguistic Task--Subjects identified which fundamental frequency was present on each trial. In the One-Dimension Condition the stimuli were /ba/-low and /ba/-high, with only fundamental frequency varying. In the Two-Dimension Condition all four stimuli occurred and subjects had to ignore variation in stop consonants. Thus there were four conditions in all: two tasks (Linguistic and Nonlinguistic) and two conditions of stimulus variation (One-Dimension and Two-Dimension). Identification times increased from One-Dimension to Two-Dimension Conditions for both tasks. However, the increase was significantly greater for the Linguistic Task. It was easier to ignore irrelevant variation in the linguistic dimension when processing the nonlinguistic dimension than vice versa.

Evidence from a variety of experimental paradigms suggests that the perception of speech and nonspeech sounds may involve processing mechanisms that are in some sense distinct. Previous approaches have generally employed a single experimental paradigm and compared the perception of speech stimuli in one condition with the perception of nonspeech stimuli in another condition. For example, in dichotic listening, speech stimuli generally yield a right-ear advantage while nonspeech stimuli generally yield a left-ear advantage (for a recent review, see Studdert-Kennedy and Shankweiler, 1970).

---

[*] Paper presented at the 82nd meeting of the Acoustical Society of America, Denver, October 1971.

[+] Haskins Laboratories and Department of Psychology, Yale University, New Haven.

[++] Neuropsychology Laboratory, Veterans Administration Hospital, West Haven, and Department of Psychology, Yale University, New Haven.

In the present experiment, we have used a different strategy. We have compared the perception of linguistic and nonlinguistic aspects of the same speech signals, by requiring subjects to identify a linguistic dimension of those stimuli in one task and a nonlinguistic dimension in another task. On each trial, a consonant-vowel syllable was presented binaurally over earphones. The subject's task was to identify which syllable had been presented. He did so by pressing one of two response buttons.

Figure 1 shows the stimuli used in both tasks. In the Linguistic Task, the stimuli were /ba/ and /da/. Both had the same (low) fundamental frequency. Whenever the subject heard /ba/-low he pressed button #1, and whenever he heard /da/-low he pressed button #2. Thus the target dimension for this task was the stop consonant, as indicated by the rectangle in Figure 1. Stop consonants were selected to represent a linguistic dimension since they appear to be the most highly encoded of all phonemes (Liberman et al., 1967).

In the Nonlinguistic Task, both stimuli were the syllable /ba/, but one had a low fundamental frequency (104 Hz) and the other had a high fundamental frequency (140 Hz). Whenever the subject heard /ba/-low he pressed button #1, and whenever he heard /ba/-high he pressed button #2. Thus the target dimension for this task was fundamental frequency, as indicated by the rectangle in Figure 1. Fundamental frequency was selected to represent a nonlinguistic dimension since it provides little or no information at the phoneme level in English.

All stimuli were prepared on the Haskins Laboratories parallel resonance synthesizer. Each was 300 msec in duration and had the same over-all intensity envelope and falling pitch contour. In the Linguistic Task the two stimuli differed only in those acoustic cues that are important for distinguishing among voiced stop consonants, namely the direction and extent of the second (Liberman et al., 1954; Delattre et al., 1955) and third (Harris et al., 1958) formant transitions. In the Nonlinguistic Task the stimuli differed only in their fundamental frequency. For each task, a block of 64 stimuli was presented in random order with an interstimulus interval of 5 sec. Each of 16 subjects received two blocks of 64 trials in each task.[1]

To summarize the two tasks: in the Linguistic Task subjects were required to identify a highly encoded linguistic dimension (stop consonants), while in the Nonlinguistic Task they were required to identify a dimension that provides little or no linguistic information at the phoneme level in English (fundamental frequency). Although we have selected stop consonants and fundamental frequency as representatives of "linguistic" and "nonlinguistic" dimensions in this experiment, we plan to examine a wide variety of other acoustic dimensions in this general paradigm. In fact, we may be able to use this approach to determine the nature of the auditory and linguistic processes which underlie the perception of any given acoustic dimension.

---

[1] For the sake of clarity, we have described the stimuli for the tasks in a simplified fashion. Each task actually used two different stimulus tapes. In the Linguistic Task, the stimuli were /ba/-low vs. /da/-low in one block, and /ba/-high vs. /da/-high in the other block. In the Nonlinguistic Task, the stimuli were /ba/-low vs. /ba/-high in one block, and /da/-low vs. /da/-high in the other block. Thus only one dimension varied in each block, and it was that dimension that the subject had to identify.

Stimuli Used in the Linguistic and Nonlinguistic Tasks

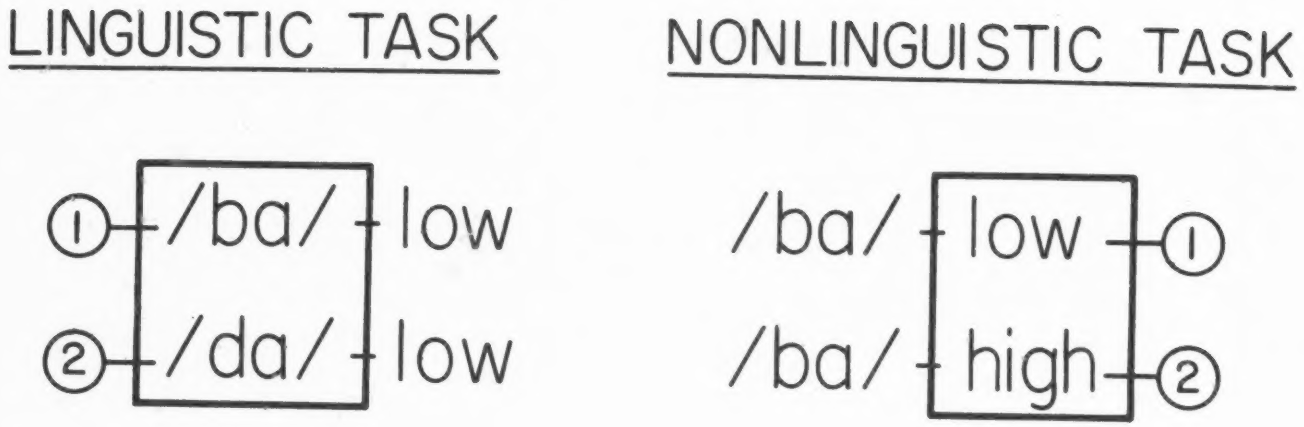


Fig. 1

Two measures of identification performance were obtained: errors and reaction time. Subjects made very few errors, less than two errors per block of 64 trials in both tasks. Therefore, error scores will not be considered further. Mean reaction times for the Linguistic and Nonlinguistic Tasks are shown in Figure 2. Each point is the mean of 128 trials for each of 16 sub- ects for a total of 2,048 trials. As shown, there was very little difference in reaction time. The 10-msec difference in reaction time was not statistically significant. Thus, under these conditions, subjects performed equally well in both tasks.

In the data just presented, only one stimulus dimension varied in each task, namely, stop consonants in the Linguistic Task and fundamental frequency in the Nonlinguistic Task. We have therefore called this condition the One-Dimension Condition. In order to examine possible interactions between the processing of linguistic and nonlinguistic information, both dimensions were varied orthogonally in another set of tests using the same subjects. The stimuli for this Two-Dimension Condition are shown in Figure 3.

In the Linguistic Task, the stimuli were /ba/-low, /ba/-high, /da/-low, nd /da/-high. As indicated by the rectangle, the target dimension in this ask was the stop consonant. Subjects pressed button #1 when they heard /ba/ nd button #2 when they heard /da/. Thus they had to ignore variations in he irrelevant dimension, fundamental frequency.

In the Nonlinguistic Task, the exact same stimuli were used: /ba/-low, ba/-high, /da/-low, and /da/-high. In this task, however, subjects identi- ied which fundamental frequency they heard on each trial. They pressed utton #1 when they heard the low fundamental frequency and button #2 when hey heard the high fundamental frequency. Thus they had to ignore variations n the irrelevant dimension, stop consonants.

What do we expect to happen in this Two-Dimension Condition relative to he One-Dimension Condition described above? If variations in the irrelevant imension significantly interfere with the processing of the target dimension e would expect reaction times to increase in the Two-Dimension Condition. eaction times for both conditions are shown in Figure 4.

The data points on the left are the same as those shown in Figure 2 for e One-Dimension Condition, in which subjects identified a given dimension en it was the only one that varied. The data points on the right are those om the Two-Dimension Condition, in which subjects performed the same identi- cation tasks but had to ignore variations in the irrelevant dimension. nsider the Linguistic Task: there was an increase in reaction time of msec from the One-Dimension to the Two-Dimension Condition. This differ- ce is highly significant. In the Nonlinguistic Task, however, the 12-msec crease in reaction time barely reached conventional levels of significance. is differential increase in reaction time in the Linguistic Task relative the Nonlinguistic Task was shown to be highly significant according to the teraction term in an analysis of variance. Thus, in the Linguistic Task, which the target dimension was stop consonants, it was very difficult to nore variations in the irrelevant dimension, fundamental frequency. How- er, in the Nonlinguistic Task, in which the target dimension was fundamental equency, subjects had very little difficulty ignoring the irrelevant dimen- on, stop consonants.

Mean Reaction Times for the Linguistic and Nonlinguistic Tasks

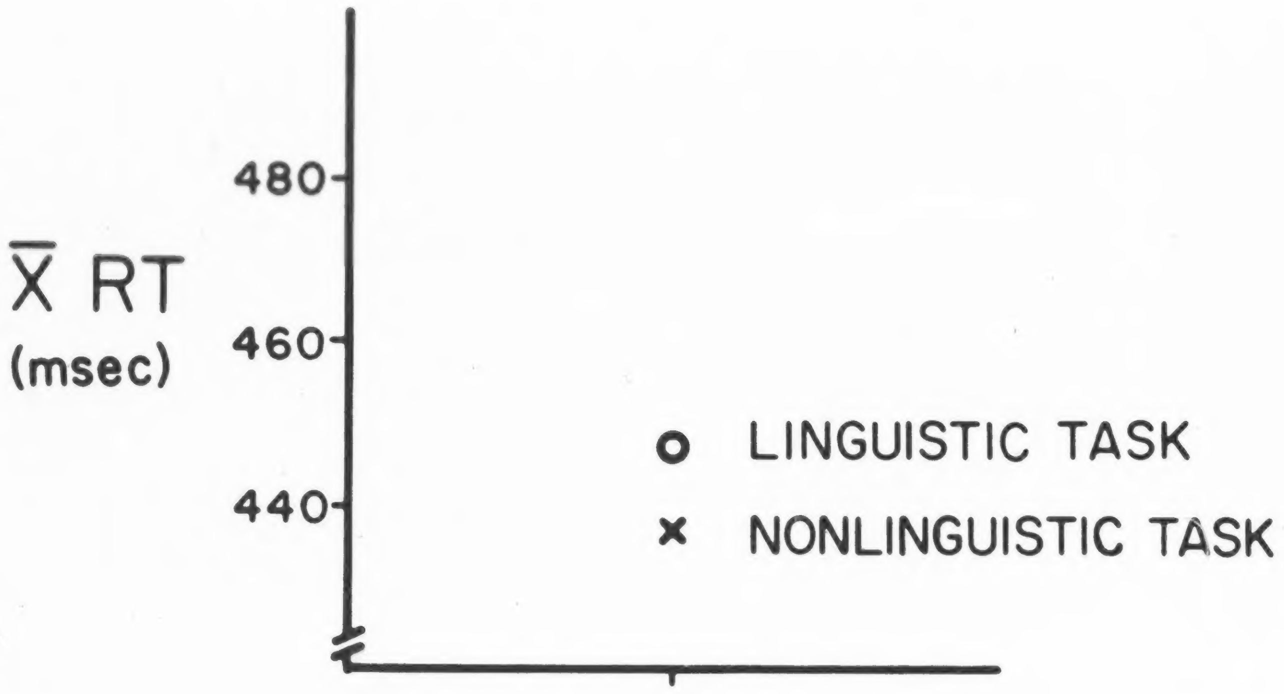


Fig. 2

Stimuli Used for the Two Tasks in the Two-Dimension Condition

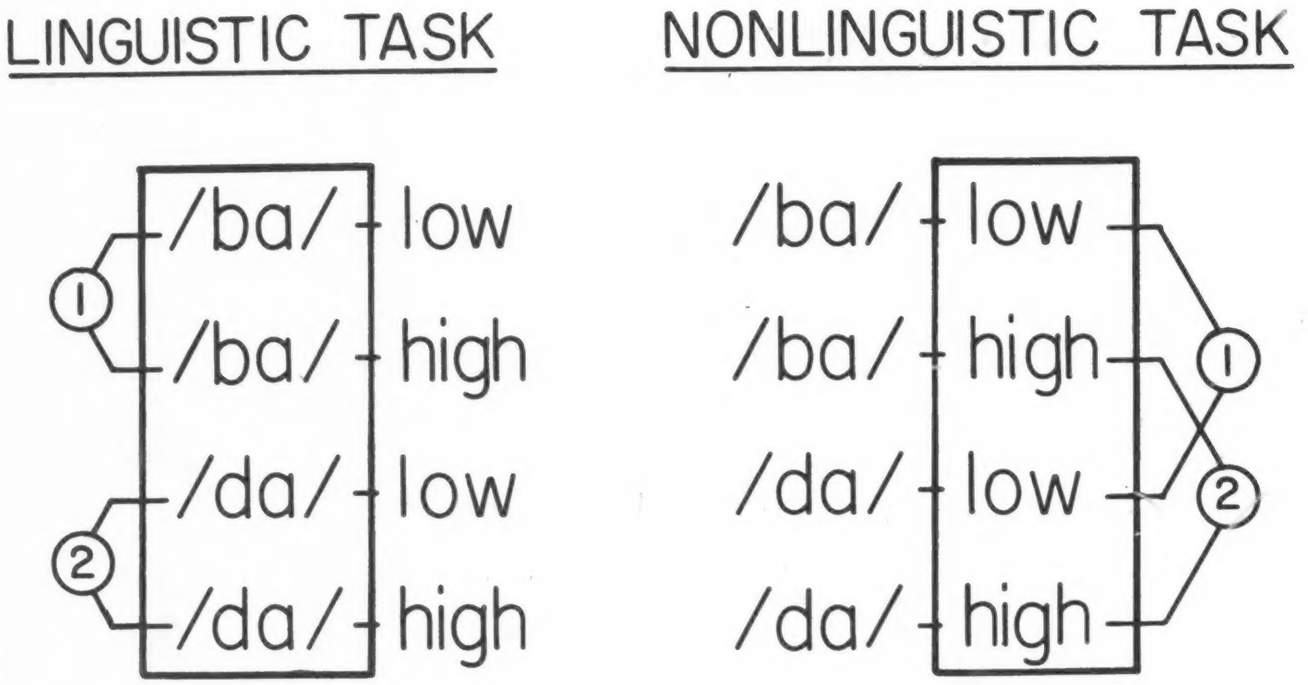


Fig. 3

Reaction Times for Both Tasks in the One-Dimension and the Two-Dimension Conditions

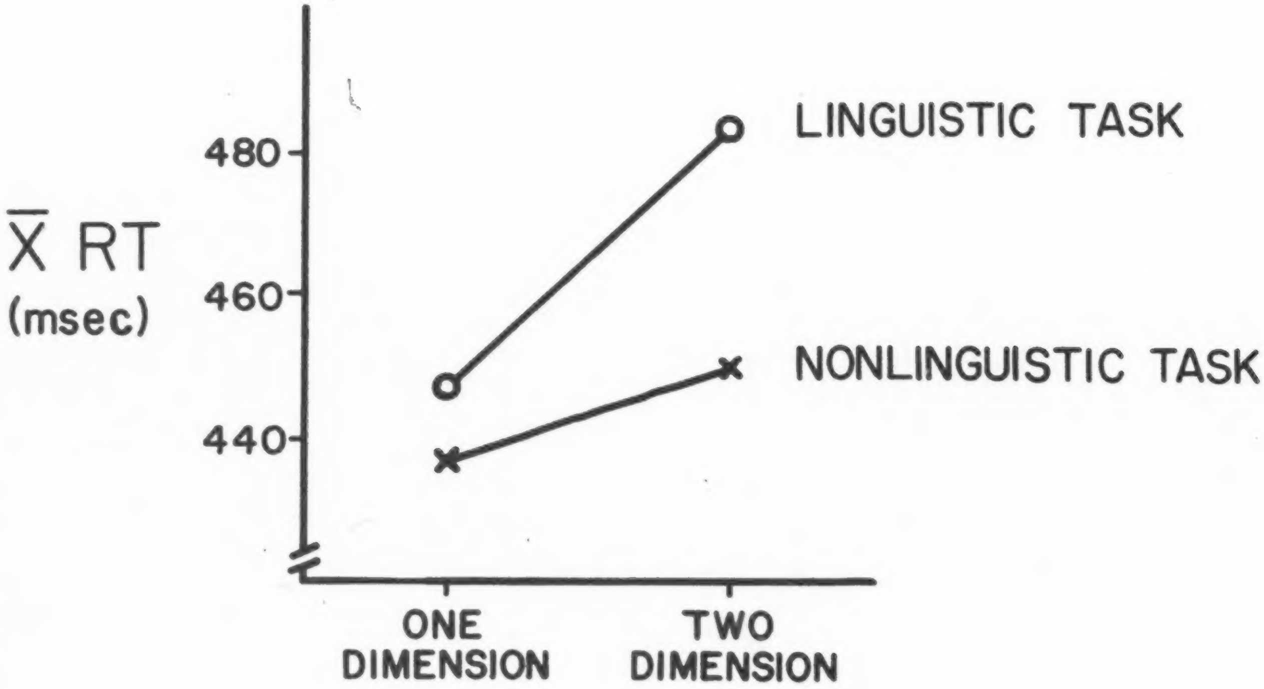


Fig. 4

To view these results in another way, examine the data from the Two-
ension Condition shown on the right in Figure 4. Recall that the stimuli
the Linguistic and Nonlinguistic Tasks were exactly the same. Therefore
only difference between tasks was which dimension subjects were required
identify. The fact that significantly different reaction times were
ained to the exact same stimuli in these tasks strongly suggests that dif-
ent perceptual processes are involved.

Elsewhere (Wood et al., 1971) we obtained averaged evoked potentials
le right-handed subjects performed the two tasks in the One-Dimension Con-
ion. We found that significantly different neural activity occurred over
left hemisphere in the Linguistic and Nonlinguistic Tasks. However, over
right hemisphere neural activity was identical in both tasks. Thus,
er the same conditions, the neurophysiological data of Wood et al. (1971)
the reaction time data of the present experiment suggest that different
eptual precesses are involved when subjects must identify linguistic vs.
inguistic dimensions.

To summarize the data of the present experiment, subjects had little
iculty ignoring stop consonants when the target dimension was fundamental
uency. In contrast, it was very difficult to ignore fundamental frequen-
hen the target dimension was stop consonants. These results suggest that
erent mechanisms underlie the processing of linguistic and nonlinguistic
nsions of the same acoustic signal.

REFERENCES

ttre, P. C., Liberman, A. M., and Cooper, F. S. (1955) Acoustic loci
and transitional cues for consonants. J. acous. Soc. Amer. 27, 769-773.
is, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., and Cooper,
F. S. (1958) Effect of third-formant transitions on the perception of
voiced stop consonants. J. acous. Soc. Amer. 30, 122-126.
rman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954)
The role of consonant-vowel transitions in the perception of the stop
and nasal consonants. Psychol. Monogr. 68.
rman, A. M., Cooper, F. S., Shankweiler, D., and Studdert-Kennedy, M.
(1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
dert-Kennedy, M. and Shankweiler, D. P. (1970) Hemispheric specialization for speech perception. J. acous. Soc. Amer. 48, 579-594.
, C. C., Goff, W. R., and Day, R. S. (1971) Auditory evoked potentials
during speech perception. Science 173, 1248-1251.

Perception of Linguistic and Nonlinguistic Dimensions of Dichotic Stimuli*

Ruth S. Day,+ James E. Cutting,+ and Paul M. Copeland+
Haskins Laboratories, New Haven

The notion that I would like to explore today is that we have different processing mechanisms for perceiving linguistic and nonlinguistic information. A major line of evidence supporting this notion comes from the dichotic listening literature. In dichotic listening a different message is presented to each ear at the same time. Typically, the subject is required to report "what he heard." Speech stimuli, such as digits, yield a right-ear advantage (Kimura, 1961). That is, subjects are more accurate in identifying stimuli presented to the right ear than those presented to the left ear. Nonspeech stimuli, such as melodies, yield a left-ear advantage (Kimura, 1964). That is, subjects are more accurate in identifying stimuli presented to the left ear.

These ear-advantage results in dichotic listening are highly replicable. How do we explain them? First, we know that language functions are handled primarily on the left side of the head. This is true for most right-handed people. An important source of evidence here is clinical: brain damage on the left side of the head usually results in language impairment, whereas comparable damage on the right side rarely interferes with language functions (for a recent review, see Geschwind, 1970).

Second, it appears that information presented to a given ear in dichotic stimulation goes primarily to the cerebral hemisphere on the opposite side of the head. Even though the ears are bilaterally represented in the two hemispheres, the pathway from a given ear to the hemisphere on the same side of the head seems to be suppressed under dichotic stimulation. Given these two assumptions (language in the left hemisphere, prepotency of crossed connections from ears to hemispheres), Kimura (1967) has explained the ear-advantage results in the following way. When both stimuli are speech, the right ear has direct access to the language-processing mechanism on the left side of the head. Meanwhile the left-ear stimulus reaches the right hemisphere and must then cross over to the left hemisphere via connecting fibers in order to undergo complete linguistic decoding. During this delay, there may be a decay in the clarity of the information, or the stimulus might undergo distortion during transmission across the connecting fibers. An analogous argument can be made for the case where both stimuli are nonspeech. The left-ear stimulus has direct access to the "nonspeech" functions of the right hemisphere, and so on. While this account is somewhat oversimplified for our purposes today, it does retain the basic features of the widely accepted Kimura model.

---

*Paper presented at meeting of the Psychonomic Society, St. Louis, November 1971.

+Also Department of Psychology, Yale University, New Haven.

Previous dichotic listening studies have retained the same experimental paradigm. They used speech stimuli in one condition and obtained a given set of results. They then used nonspeech stimuli in another condition and obtained a contrasting set of results. In the work I will discuss today, we have used a very different strategy. We have used only speech stimuli. But we have required subjects to track a linguistic dimension in one condition, and a nonlinguistic dimension of the same stimuli in another condition.

Method

Stimuli. The stimuli were the consonant-vowel (CV) syllables /bæ, dæ, gæ/. Each had three pitch levels: high, medium, and low fundamental frequency. Thus there were nine stimuli in all: /bæ/-high, /bæ/-medium, /bæ/-low, /dæ/-high/ dæ/-medium, /dæ/-low, /gæ/-high, /gæ/-medium, /gæ/-low. They were synthesized on the parallel resonance synthesizer at the Haskins Laboratories. All syllables were 300 msec in duration and had identical intensity envelopes.

The /bæ/'s, /dæ/'s, and /gæ/'s differed from each other only in those cues known to be important for discriminating among voiced stop consonants. These cues are the direction and extent of the second (Liberman et al., 1954; Delattre et al., 1955) and third formant transitions (Harris et al., 1958). Stop consonants were selected to represent the linguistic dimension since they are the most highly encoded of all speech sounds (Liberman et al., 1967).

The highs, mediums, and lows differed only in their fundamental frequency. Each had a falling pitch contour, but began and ended at nonoverlapping frequency values. They began at 166, 130, and 96 Hz, respectively, and each fell 10 Hz. Fundamental frequency was selected to represent the nonlinguistic dimension since it provides little or no linguistic information at the phoneme level in English.

To summarize: the nine stimuli were classifiable according to two dimensions: a linguistic dimension (stop consonants) and a nonlinguistic dimension (fundamental frequency). Both dimensions were highly discriminable, as shown by the appropriate pre-tests.

Tapes. Dichotic tapes were prepared on the pulse code modulation system at Haskins. This system enables the experimenter to line up the onsets of dichotic stimuli with an accuracy of 1/2 msec. The stimulus pairs were varied in relative onset time. Sometimes the left-ear stimulus began first by 50 msec; on other trials the right-ear stimulus began first by 50 msec; and on remaining trials both stimuli began at the same time.

Procedures. The subject's task was to determine which stimulus began first on each trial. Thus, he had to make a temporal order judgment (TOJ). There were two conditions. 1) Linguistic Task: subjects had to report which stop consonant began first, /b/, /d/, or /g/. 2) Nonlinguistic Task: subjects had to report the pitch level of the leading stimulus, high, medium, or low. The same stimulus tapes were used for both conditions. All 16 subjects performed both tasks. They were right-handed, native English speakers and had no history of hearing trouble. All the appropriate counterbalancing procedures were observed, with respect both to test order and to the arrangement of items on the tape.

Percent Correct Temporal Order Judgment (TOJ)
for the Linguistic and Nonlinguistic Tasks

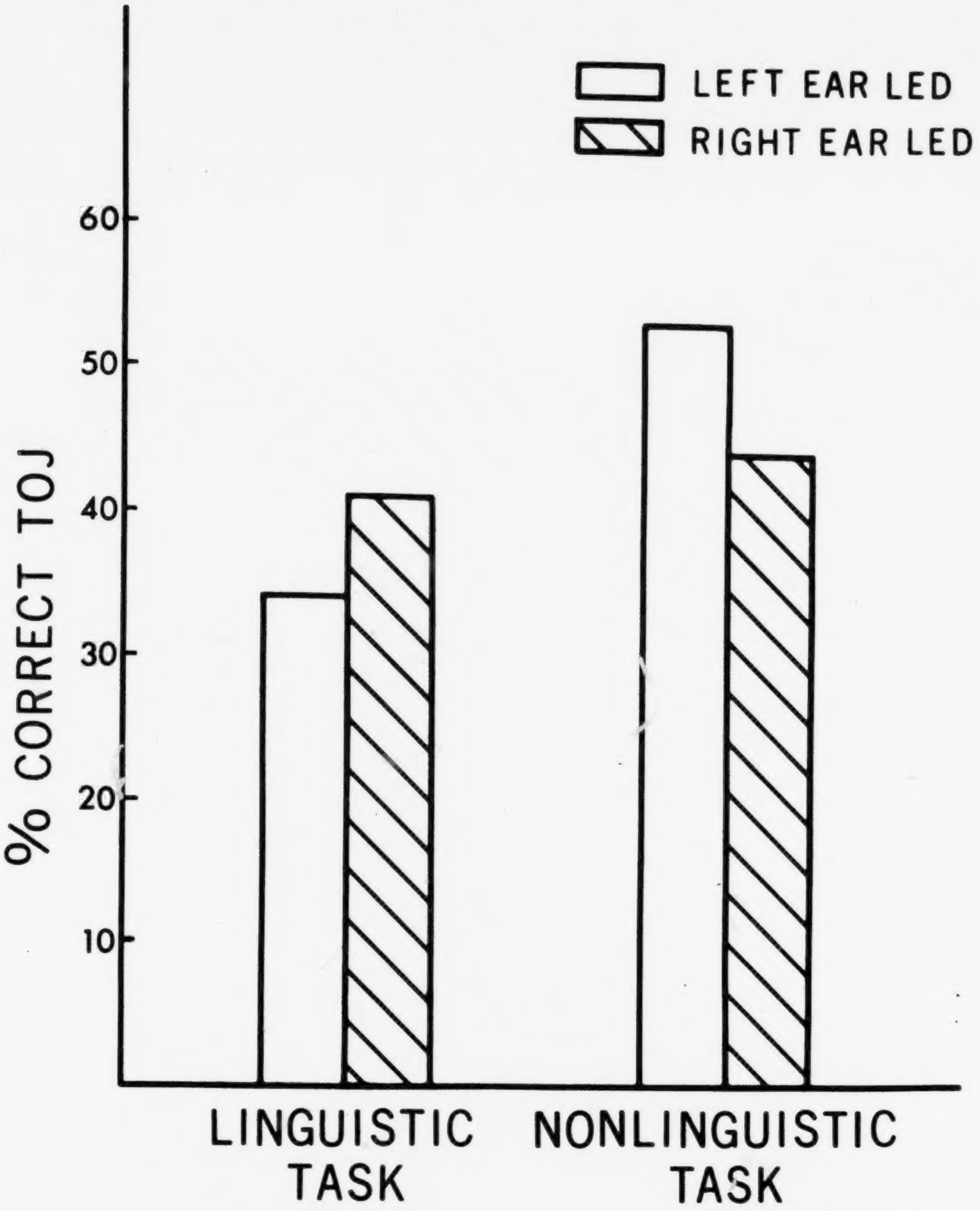


Fig. 1

Results and Discussion

Linguistic Task. When subjects had to determine which stop consonant led, there was a right-ear advantage. That is, on those trials where the right-ear stimulus led, subjects were 41% correct in judging temporal order; on those trials where the left-ear stimulus led, they were 34% correct. Thus there was a 7% net advantage in favor of the right ear.

Nonlinguistic Task. When subjects had to determine the pitch on those trials where the left-ear stimulus led, subjects were 53% correct, while they were only 44% correct when the right-ear stimulus led. Thus there was a net 9% advantage in favor of the left ear. The results for both conditions are summarized in Figure 1.

Note that we have used the same stimuli in both tasks. Therefore the ear advantages could not have been determined by the nature of the stimuli. Instead it was the nature of the task requirements that determined the direction of the ear advantage: when subjects had to target for the nonlinguistic dimension of the same stimuli there was a left-ear advantage. These results are compatible with those of previous dichotic listening studies that used speech and nonspeech stimuli in separate conditions. Yet they go on to suggest that different processing mechanisms are involved in tracking linguistic and nonlinguistic dimensions of the same acoustic stimuli.

Despite the differences in ear advantage between the two tasks, the effect was not statistically significant. Perhaps the presence of variation in the irrelevant dimension attenuated the magnitude of these ear-advantage results. In order to study this possibility, we are currently retesting the same subjects. Again they judge the temporal order of stops in the Linguistic Task and fundamental frequency in the Nonlinguistic Task. However, the target dimension is the only one that varies. Hopefully, the ear-advantage data will be more sizeable in this situation.

There is another way to look at the ear-advantage data of the present experiment. Given that a subject had a particular value of an ear advantage on the Linguistic Task, did his score move "leftward" on the Nonlinguistic Task? The answer is yes: 12 subjects moved leftward, 3 moved rightward, and 1 showed no change. This shift in ear advantage was significant as shown by the Task x Ear interaction term in an analysis of variance ($F = 4.76$, $p < .05$).

There was another finding of considerable interest. Let's put the whole issue of ear advantage aside. Instead, consider over-all performance levels for the two tasks. Performance was better on the Nonlinguistic Task (49% correct) than on the Linguistic Task (38% correct) ($F = 13.27$, $p < .005$). This is what we would expect if an additional processor is needed in order to decode linguistic information. Both tasks require judgment of temporal order. However, the stimuli in the Linguistic Task may require more complicated analysis than do these same stimuli in the Nonlinguistic Task. These task differences support the notion that a special decoder is needed to handle linguistic information.

The specialized decoder notion receives additional support from some recent experiments in which we used a different experimental paradigm. Each trial consisted of a single binaural stimulus that subjects had to identify.

In the Linguistic Task they had to identify which stop consonant had occurred, while in the Nonlinguistic Task they had to identify which fundamental frequency had occurred. Our strategy was the same as in the present experiment: we used the same acoustic stimuli but required subjects to track different dimensions of these stimuli in the two tasks. Again, we obtained task differences, this time in terms of reaction time (Day and Wood, 1971) and neural activity (Wood, Goff, and Day, 1971).

To summarize the present experiment: subjects judged the temporal order of dichotic stimuli that varied along a linguistic and a nonlinguistic dimension. When subjects had to target for the linguistic dimension, there was a right-ear advantage. When they had to target for the nonlinguistic dimension there was a left-ear advantage. This shift in ear advantage between the two tasks was significant. Finally, over-all performance was better on the Nonlinguistic Task. These results, collectively, suggest that there are different processing mechanisms for linguistic and nonlinguistic[1] information.

REFERENCES

Day, R.S. and Wood, C.C. (1971) Interactions between linguistic and nonlinguistic processing. (See this Status Report.)

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955) Acoustic loci and transitional cues for consonants. J. acous. Soc. Amer., 27, 769-773.

Geschwind, N. (1970) The organization of language and the brain. Science 170, 940-944.

Harris, K.S., Hoffman, H.S., Liberman, A.M., Delattre, P.C., and Cooper, F.S. (1958) Effect of third-formant transitions on the perception of voiced stop consonants. J. acous. Soc. Amer. 30, 122-126.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. exp. Psychol. 16, 355-358.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. 68.

Liberman, A.M., Cooper, F.S., Shankweiler, D., and Studdert-Kennedy, M. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Wood, C.C., Goff, W.R., and Day, R.S. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

---

[1] We plan to extend our study to native speakers of tone languages such as Thai since pitch level is a linguistic dimension in these languages.

Dichotic Backward Masking of Complex Sounds*

C.J. Darwin[+]
Haskins Laboratories, New Haven

ABSTRACT

In the first experiment subjects identified a consonant-vowel syllable presented dichotically with a known masking sound at a stimulus onset asynchrony of ± 60 msec. When the mask followed the target syllable, perception of place of articulation of the consonant was impaired more when the mask was a different consonant-vowel syllable than when it was either a steady-state vowel or a nonspeech timbre. Perception was disturbed less when the mask preceded the target, and the amount of disruption was independent of which mask was used. Greater backward than forward masking was also found in the second experiment for the identification of complex sounds which differed in an initial change in pitch. These experiments suggest that the extraction of complex auditory features from a target can be disrupted by the subsequent dichotic presentation of a sound sharing certain features with the target.

The traditional task in experiments on the temporal course of auditory masking has been the detection of a target presented in close temporal proximity to a mask. This paradigm has shown only small effects when target and mask are presented to opposite ears (dichotically). Moreover, these effects have been found only over very brief stimulus-mask intervals. Elliott (1962), for example, found virtually no forward masking of a brief tone by contralateral white noise, and only slight backward masking extending out to an interstimulus interval of about 15 msec.

Recently Studdert-Kennedy, Shankweiler, and Schulman (1970) have reported an experiment requiring identification of two stop-consonant syllables presented dichotically with a temporal offset between them. They found that for offsets between about 15 and 120 msec the lagging syllable was reported more accurately than the leading syllable. Their result has since been confirmed both in the original paradigm (Berlin et al., 1970; Lowe et al., 1970) and in a slightly different paradigm in which only one sound has to be reported on a single trial (Kirstein, 1970; 1971). No advantage for the lagging over the leading sound, however, was found in binaural presentation (with both syllables coming to both ears) even when the duration of the vowel portion of each syllable was drastically curtailed to eliminate temporal overlap between the two sounds (Porter, 1971). Such curtailing did not influence the dichotic effect.

---

* Paper to appear in The Quarterly Journal of Experimental Psychology, 23, Part 4 (November 1971).

[+] Now at the University of Sussex, England.

In terms of masking, these experiments have shown that under dichotic presentation stop-vowel syllables are more effective as mutual backward than as mutual forward maskers, whereas under binaural presentation, provided they do not temporally overlap, any masking that occurs is symmetrical.

In the visual modality, dichoptic masking is essentially a contour interaction (Schiller, 1965; Kahneman, 1968), which is asymmetrical so that backward masking is greater than forward. This asymmetry supports theories which emphasize the interruption of perceptual processes by the mask, rather than a temporal summation of mask and target (Kahneman, 1968; Spencer and Shuntich, 1970). A similar explanation seems appropriate for the auditory case (Studdert-Kennedy et al., 1970), although for stop-vowel syllables the effect is confined to dichotic presentation, whereas in vision monoptic contour interactions can be obtained (Schiller, 1965).

The present study pursues the analogy between dichoptic and dichotic masking. In the auditory experiments reviewed above it is not clear whether the superior backward over forward masking is confined to a particular type of mask, since only syllables have been used to mask syllables. The first experiment examines the relative extent of forward and backward masking for a number of different masks on a stop-vowel target set.

EXPERIMENT I

The masks used in this experiment were chosen to have certain properties in common with the target set. Three were speech and the fourth a nonspeech timbre. The three speech sounds were (1) a steady-state vowel different from that used in the target syllables, (2) the same vowel as used in the target syllables, and (3) a stop-vowel syllable with the same vowel as the targets but a different stop consonant.

Method

The targets used in this experiment were the four stop-vowel syllables /bɛ , dɛ, pɛ, tɛ /. These four consonants give two values each on the traditional phonetic dimensions of place of articulation and voicing. The four masks were /gɛ, ɛ, ɔ/ and a nonspeech steady-state timbre, which had three formants at 894, 2910, and 3698 Hz, respectively. The two steady-state vowel masks and the five syllables were all highly intelligible. All the sounds were synthesized with three formants on the Haskins Parallel Formant Synthesizer. Each sound lasted 100 msec, and all the sounds had the same intonation contour and were equated for peak amplitude on a VU meter. On each trial of the experiment a subject heard one of the target sounds in one ear and one of the masks in the other. He always knew which mask would occur since this was held constant over a block of forty-eight trials and was played to him before each block, but he did not know into which ear the target would come. His task was simply to identify which of the four targets had been presented; he did not have to say into which ear it had come. The sounds on the two ears were always temporally offset by 60 msec. Whether the target or the mask led was randomly determined with the restriction that within each block of forty-eight trials each target item led six times and lagged six times. Sixteen subjects each took eight blocks of forty-eight trials in a Latin square design which counterbalanced the order in which the four masks were heard. The subjects were given a binaural demonstration of the set of target items before taking the dichotic test. Before each block the mask for that block was played three times binaurally.

Results

Three different scoring methods were used: (1) the response had to be entirely correct (both place of articulation and voicing), (2) only place of articulation had to be correct, and (3) only voicing had to be correct. The results according to these three methods are shown in Figure 1. The slope of each line indicates the difference between the target leading and target lagging conditions for the various masks. A line with a positive slope indicates that the target is better perceived when it lags the mask than when it leads it.

Look first at the results where both place of articulation and voicing had to be correct. Analysis of variance on this data showed a significant interaction of the lead/lag factor with mask [$F(3,105) = 3.98$; $p < .01$]. However, since an analysis of variance on the results for place of articulation and voicing separately showed a significant difference between these two features for the interaction of lead/lag and mask [$F(3,45) = 3.16$; $p < .05$] as well as a significant interaction between the feature analyzed and lead/lag condition [$F(1,15) = 23.8$; $p < .001$], the results will now be described separately for these two features.

For place of articulation, as with both features combined, there was a significant interaction of mask with whether the target led or lagged the mask [$F(3,45) = 12.5$; $p < .001$]. However, as is clear from the figure, this interaction is mainly due to the case when the target leads the mask (i.e., to the backward masking case). This was confirmed in analysis of variance which showed a highly significant effect of mask on a preceding target [$F(3,45) = 18.6$; $p < .001$], but only slight variation when the target follows the mask [$F(3,45) = 2.75$; $.1 > p > .05$]. Thus for perception of place of articulation the amount of forward masking is virtually independent of the mask, but the amount of backward masking is much greater when the mask is another stop-vowel syllable than when it is one of the other masks ($p < .001$). However, the three steady-state masks do show significantly greater backward than forward masking ($p < .001$) although the amount of backward masking is very much less than for /gɛ/.

For the extraction of voicing, however, there was no overall advantage for the lagging over the leading target ($F < 1.0$) and only a slight interaction of lead/lag condition with mask [$F(3,45) = 2.46$; $.1 > p > .05$]. Thus the perception of voicing shows no more backward than forward masking for the masks used here.

In summary this experiment shows that for stop-vowel syllables dichotically opposed by a mask at temporal offsets of ± 60 msec: (1) forward masking is roughly constant for the four masks used, for both place of articulation and voicing; (2) backward masking is greater than forward for place of articulation but not for voicing for all the masks; but (3) this difference is considerably greater when the mask is another stop-vowel syllable than when it is the same vowel, a different vowel, or a nonspeech timbre; (4) these last three masks do not differ significantly in any condition.

Mean Percents Correct for Stop-Vowel Syllables Dichotically Opposed by a Mask at ± 60 Msec Offset

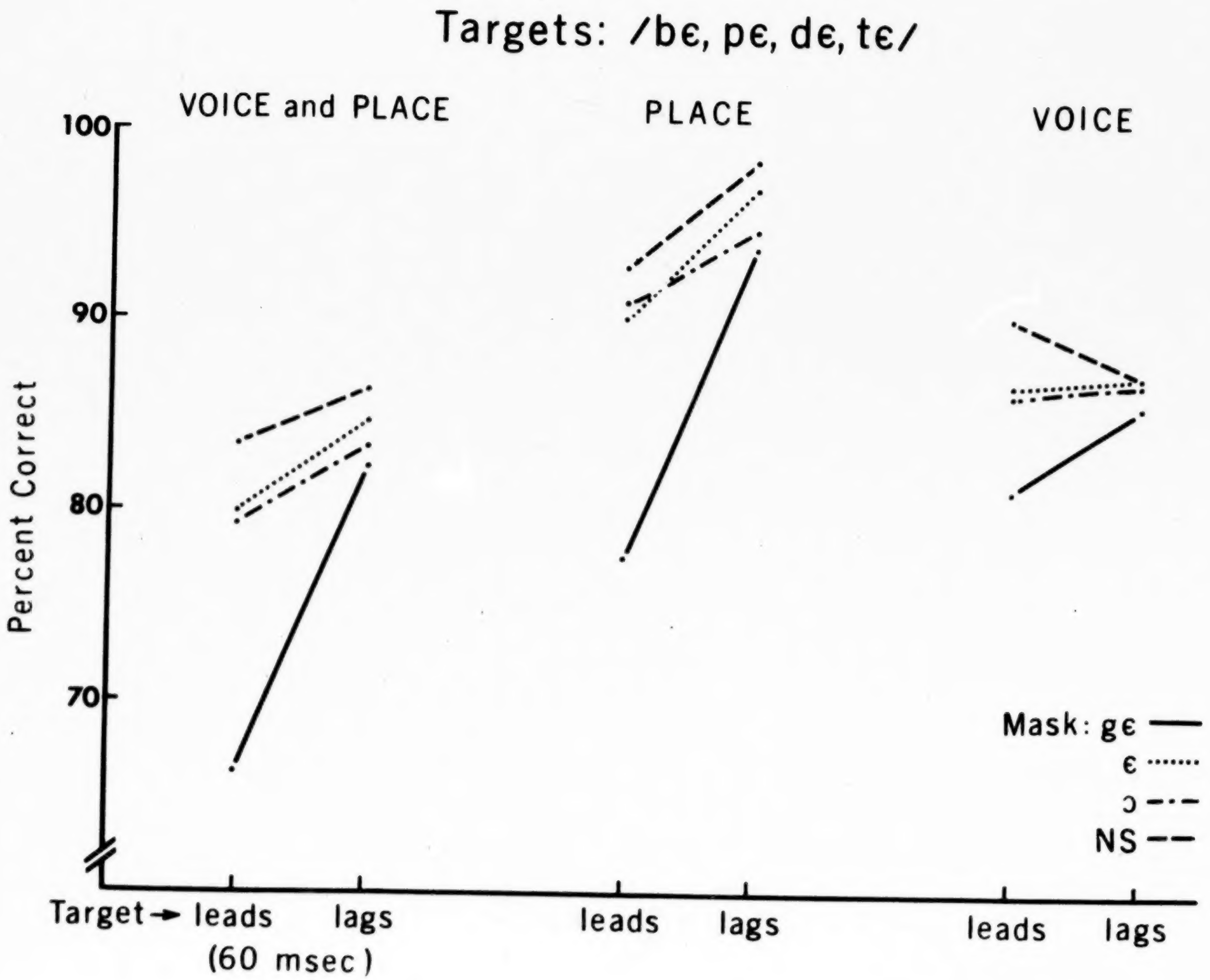


Fig. 1

The three columns refer to three different scoring criteria.

Discussion

The amount of backward masking, at least for the perception of place of articulation, is clearly dependent on the mask used. Dichotic masking is thus a potentially useful tool for describing features in auditory perception. The sharp discontinuity between the effects of the three steady-state masks and the /gɛ/ mask argues against any general continuum of similarity being important, for if it were we might have expected the /ɛ/ mask to have been closer in its effect to the /gɛ/ mask. Rather, we are led to suppose that the /gɛ/ mask contains specific features which are particularly effective at interrupting the perception of the preceding target. This interpretation is strengthened by the absence of any mask specificity in the forward masking case, although this may be at least partly due to the very high performance leaving little room for improvement.

Two more points require discussion: the slight, though consistently greater, effect of backward over forward masking for the three steady-state masks and the absence of any differences either between masks or between the forward and backward conditions for the perception of voicing. The first point may be attributable to some unspecific auditory effect or perhaps may not even be specific to the auditory modality. A kick on the shins may be an effective backward mask to this extent. The effect is quite small and will probably be difficult to investigate. The absence of any interesting effects in the perception of voicing may reflect the very different acoustic cues underlying the perception of voicing and of place of articulation. For voicing, at least in this experiment, the detection of some aspiration at the beginning of the stimulus would give sufficient information, whereas for place of articulation detailed knowledge of the slope of rapid formant transitions is required. The extraction of this latter information may be particularly sensitive to disruption.

This experiment alone cannot decide whether extraction of the acoustic parameters, on which the decision concerning place of articulation is based, is being disturbed or whether it is rather some purely linguistic process such as the relation of these acoustic features to a linguistic framework. To distinguish between these two hypotheses the next experiment looks at backward masking for stimuli which, like stop-vowel syllables, are distinguished by a rapidly changing initial section, but which are not perceived as falling into different phonemic categories.

EXPERIMENT II

This experiment uses a paradigm introduced by Kirstein (1970). No *a priori* distinction is made between target and mask, both being drawn from the same stimulus set. The subject attends to one ear and is asked to recall the stimulus presented there.

Method

Three different sounds were used. They differed in their fundamental frequency contours which are illustrated in Figure 2. These pitch contours were carried on the steady vowel /ɛ/. Dichotic pairs were made up using the Haskins Parallel Formant Synthesizer and a special computer program (Mattingly, 1968)

Pitch Contours of the Three Sounds Used in Experiment II

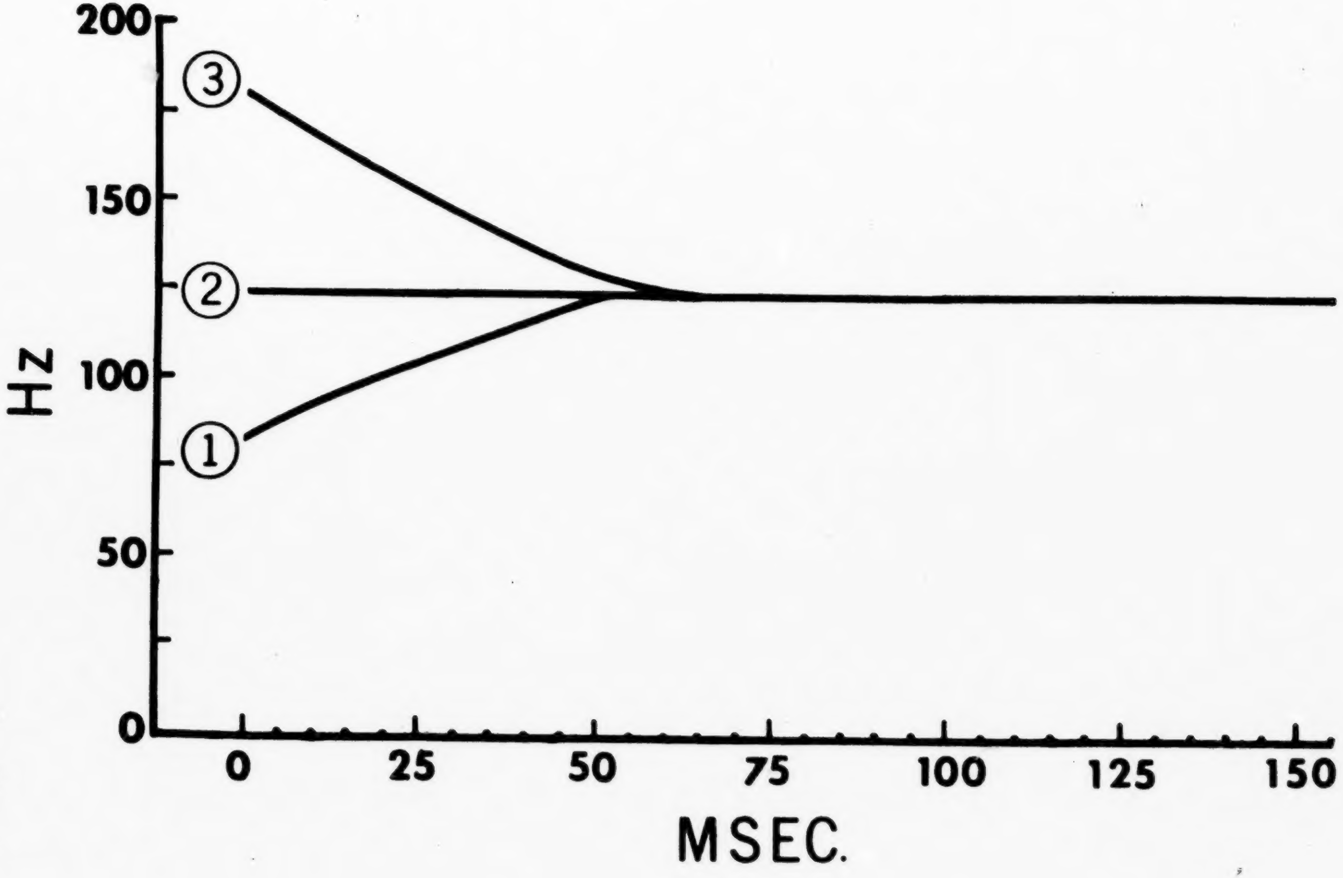


Fig. 2

which ensured perfect timing of the signals on the tape's two tracks. On each trial a subject heard two pitch contours, one in either ear. They were either simultaneous or offset by ± 25 msec. Subjects attended to one ear for a block of trials and were instructed to identify only the sound that was presented to that ear. They were given training in identifying the three sounds with the first three digits. Half the pairs of sounds they heard were simultaneous and half were temporally offset. Twelve right-handed subjects took the experiment in a procedure which counterbalanced ears and attention.

TABLE 1

Mean Percents Correct in Experiment II According to Asynchrony of the Reported Stimulus

	Asynchrony of Reported Stimulus					
	Simultaneous		Leading		Lagging	
	left	right	left	right	left	right
Attend left	40.4	37.1	35.2	33.0	48.6	43.6
Attend right	36.6	40.4	29.5	37.8	40.9	48.6
Total	38.5	38.7	32.4	35.4	44.8	46.1

Results

The results are tabulated (Table 1) in terms of the asynchrony of the reported stimulus. Thus, if the subject were presented with stimulus 1 to his left ear 25 msec ahead of stimulus 2 to his right and, though asked to report the left ear, in fact wrote "2," a correct response would have been recorded for the right-ear-lagging--attend-left cell. There was a clear advantage for the lagging over the leading condition irrespective of ear or attention condition (twelve subjects for, none against). Subjects were generally poor at selecting the requested ear though there was some indication that selective attention was easier for the staggered pairs than for the simultaneous ($p < .1$). There was no difference between the ears in either the simultaneous or the staggered condition ($p > .1$).

Discussion

As in the first experiment we find here greater backward than forward dichotic masking for sounds which are distinguished by a rapidly changing initial portion. In the first experiment rapidly changing formant transitions cued the place of articulation distinction in stops, whereas in this experiment the sounds have been distinguished by changes in fundamental frequency which do not cue a phonemic distinction. Parsimony suggests that explanations for these effects should be sought at a purely auditory level of analysis rather than supposing that separate explanations are required for both speech and nonspeech sounds.

Two brief comments on subsidiary results of the second experiment. First, the slightly more efficient selective attention under staggered than under simultaneous conditions bears out a suggestion by Treisman and Riley (1969) to that effect. Second, the absence of any ear difference here contrasts in an interesting way with Haggard and Parkinson's (1971) finding that when rapid pitch changes similar to the ones used here cue the voiced/voiceless distinction in stops (Haggard et al., 1970) there is an advantage for the right ear under dichotic presentation. The right-ear advantage is thus determined by the use to which acoustic information is put rather than by the presence of particular acoustic features (Darwin, 1971; Haggard and Parkinson, 1971).

Both experiments reported here have used dichotic presentation. As mentioned in the introduction, the superiority of backward over forward masking for stop-vowel syllables by similar syllables is not found for monaural presentation (Porter, 1971). A related finding is that by Massaro (1970), who finds slightly larger dichotic than monotic backward masking for the identification of a pure tone followed by a longer masking tone. This greater effectiveness for dichotic over monotic presentation may reflect a segmentation problem faced by the auditory system in determining whether interruption of the preceding signal should occur or whether a second signal should be treated as part of the same perceptual sequence. Misplaced interruptions would be restricted to a minimum, at least in natural situations, if sounds from the same source were treated without interruption. Spatial location could provide a very reliable criterion for determining whether temporally distinct sounds originated from a common source. Spatial location has the added advantage that at least its directional aspect has neurophysiological correlates at a very peripheral level of the auditory system. This information is thus potentially available for guiding the sequential analysis of the auditory input at higher levels. This may not be the only criterion, and indeed Massaro's experiments with pure tones show appreciable monotic backward masking. But different processes may be operating for simple and complex stimuli, since Massaro also finds backward masking for pure tones to be relatively independent of the similarity of test and masking tones.

REFERENCES

Berlin, C.I., Willett, M.E., Thompson, C., Cullen, J.K. and Lowe, S.S. (1970) Voiceless versus voiced CV perception in dichotic and monotic listening. J. acoust. Soc. Amer. 47, 75(A).

Darwin, C.J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. exp. Psychol. 23, 46-62.

Elliott, L.L. (1962) Backward masking: Monotic and dichotic. J. acoust. Soc. Amer. 34, 1108-1115.

Haggard, M.P. and Parkinson, A.M. (1971) Stimulus and task factors as determinants of ear advantages. Quart. J. exp. Psychol. 23, 168-171.

Haggard, M.P., Ambler, S. and Callow, M. (1970) Pitch as a voicing cue. J. acoust. Soc. Amer. 47, 613-617.

Kahneman, D. (1968) Method findings and theory in studies of visual masking. Psych. Bull. 70, 404-425.

Kirstein, E.F. (1970) Selective listening for temporally staggered dichotic CV syllables. J. acoust. Soc. Amer. 48, 95(A).

Kirstein, E.F. (1971) Temporal factors in perception of dichotically presented stop consonants and vowels. Ph.D. dissertation, Univ. of Connecticut. (Haskins Laboratories Status Report 24, Oct.-Dec. 1970).

Lowe, S.S., Cullen, J.K., Thompson, C., Berlin, C.I., Kirkpatrick, L.L., and Ryan, J.T. (1970) Dichotic and monotic simultaneous and time-staggered speech. J. acoust. Soc. Amer. 47, 76(A).

Massaro, D.M. (1970) Preperceptual auditory images. J. exp. Psychol. 85, 411-417.

Mattingly, I.G. (1968) Experimental methods for speech synthesis by rule. IEEE Trans. AU-16, 198-202.

Porter, R.J. (1971) The effect of temporal overlap on the perception of dichotically and monotically presented CV syllables. Paper presented at the 81st Meeting of Acoustical Society of America, Washington, D.C.

Schiller, P.H. (1965) Monoptic and dichoptic visual masking by patterns and flashes. J. exp. Psychol. 69, 193-199.

Spencer, T.J. and Shuntich, R. (1970) Evidence for an interruption theory of backward masking. J. exp. Psychol. 85, 198-203.

Studdert-Kennedy, M., Shankweiler, D.P. and Schulman, S. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. acoust. Soc. Amer. 48, 599-602.

Treisman, A.M. and Riley, J.G.A. (1969) Is selective attention selective perception or selective response? A further test. J. exp. Psychol. 79, 27-34.

ABSTRACT

On the Nature of Categorical Perception of Speech Sounds[*]

David Bob Pisoni[+]
Haskins Laboratories, New Haven

Current theories of speech perception emphasize that the perception of speech sounds may involve processes that are in some way basically different from the processes involved in the perception of other sounds. One of the findings which has been cited as evidence for a special mode of perception is the differences in perception between synthetic stop consonants and steady-state vowels. Stop consonants have been found to be perceived in a categorical mode, unlike other auditory stimuli. Discrimination is limited by absolute identification. Listeners are able to discriminate stimuli drawn from different phonetic categories but cannot discriminate stimuli drawn from the same phonetic category, even though the acoustic distance between stimuli is comparable. On the other hand, steady-state vowels have been found to be perceived continuously. Discrimination is independent of category assignment. Listeners are able to discriminate many more differences than would be predicted on the basis of absolute identification.

The primary goal of the present investigation was to examine the differences between categorical and continuous perception and to evaluate three different explanations for the phenomena of categorical perception. Six experiments dealing with the identification and discrimination of synthetic speech sounds were conducted to determine the nature of categorical perception. The first experiment replicated the original findings on the differences in perception between consonants and vowels reported by investigators at Haskins Laboratories. Perception of stop consonants was found to be "nearly categorical" in the sense that listeners tend to discriminate pairs of stimuli only to the extent that they identify them as different. Perception of steady-state vowels was found to be more "nearly continuous" in the sense that the same listeners discriminate many more intraphonemic differences than they identify absolutely.

The second experiment attempted to assess the effects of discrimination training with non-speech stimuli on categorical perception. The results indicated that there were large individual differences among _Ss_ and that no definite conclusions could be drawn about the effects of discrimination training in producing categorical perception with non-speech stimuli.

The third experiment considered an explanation of categorical perception in terms of the auditory and phonetic processes involved in speech discrimination tasks. It was found that steady-state vowels tend to be perceived more categorically at brief stimulus durations. The results also confirmed predictions derived

---

[*] Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The University of Michigan.

[+] Now at Indiana University, Bloomington.

from a model proposed by Fujisaki which suggested that auditory short-term memory may be involved in speech discrimination.

The fourth and fifth experiments were concerned with comparing a new discrimination procedure, the four-interval test of paired similarity (4IAX), with the traditional ABX discrimination test. It was found that substantial differences in discrimination may be obtained with the 4IAX procedure as compared with the ABX for vowels, while less marked differences in discrimination may be obtained with consonants.

The sixth experiment tested the hypothesis that consonants and vowels differ in the degree to which auditory short-term memory is employed in their discrimination. The results of a delayed comparison recognition memory task indicated that accuracy of discrimination for vowels both within and between phoneme boundaries was related to the magnitude of the comparison interval. In contrast, discrimination of stop consonants remained relatively stable both within and between phoneme boundaries.

The results of this investigation suggested that the major differences between categorically and continuously perceived speech stimuli are related to the differential availability of auditory short-term memory for the acoustic cues distinguishing different classes of speech sounds. For highly encoded speech sounds such as stop consonants, within-category discrimination is so poor as to suggest that information other than a binding phonetic categorization is unavailable to the listener for use in discrimination.

ERRATUM

Haskins Laboratories Status Report on Speech Research, SR-24 (1971)

Letter Confusions and Reversals of Sequence in the Beginning Reader: Implications for Orton's Theory of Developmental Dyslexia.

Isabelle Y. Liberman, Donald Shankweiler, Charles Orlando, Katherine S. Harris, and Fredericka B. Berti.

A computational error requires correction of Table III and minor changes in the text. The error led to underestimation of the error rate for other consonants (OC) because the number of opportunities for error had been wrongly calculated. In the column of Table III headed "Other Consonant," the opportunities should be 2736 and the percent should be 16.3.

In the text, two changes are required. (1) On page 25, the first sentence should be amended to read as follows: "Reversals of orientation (RO) have a greater relative frequency of occurrence than sequence reversals (RS), but less than other consonant errors (OC).[3]" (2) On page 27 in the first paragraph of the discussion, the third sentence should be changed to read as follows: "Viewed in terms of opportunities for error, RO's occurred less frequently than other consonant errors."

PUBLICATIONS AND REPORTS*

Publications and Manuscripts

Observing Laryngeal Adjustments During Running Speech by Use of a Fiberoptics System. M. Sawashima, A.S. Abramson, F.S. Cooper, and L. Lisker. Phonetica (1970) 22, 193-201.

The Motor Theory of Speech Perception: A Reply to Lane's Critical Review. M. Studdert-Kennedy, A.M. Liberman, K.S. Harris, and F.S. Cooper. Psychological Review (1970) 77, 234-249.

Supraglottal Air Pressure in the Production of English Stops. L. Lisker. Language and Speech (1970) 13, 215-230.

Towards a Unified Phonetic Theory. P. Lieberman. Linguistic Inquiry (1970) 1, 307-322.

Discrimination in Speech and Nonspeech Modes. I.G. Mattingly, A.M. Liberman, A.K. Syrdal, and T. Halwes. Cognitive Psychology (1971) 2, 131-157.

On the Speech of Neanderthal Man. P. Lieberman and E.S. Crelin. Linguistic Inquiry (1971) 2, 203-222.

Auditory Evoked Potentials During Speech Perception. C.C. Wood, W.R. Goff, and R.S. Day. Science (1971) 173, 1248-1251.

Background of the Conference. A.M. Liberman and J.J. Jenkins. Introduction to the Conference on Communicating by Language: The Relationships between Speech and Learning to Read, Elkridge, Md., May 1971. To appear in Language by Ear and by Eye: The Relationships between Speech and Reading, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press, in press).

How is Language Conveyed by Speech? F.S. Cooper. Paper presented at the Conference on Communicating by Language: The Relationships between Speech and Learning to Read, Elkridge, Md., May 1971. To appear in Language by Ear and by Eye: The Relationships between Speech and Reading, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press, in press).

Reading, the Linguistic Process, and Linguistic Awareness. I.G. Mattingly. Paper presented at the Conference on Communicating by Language: The Relationships between Speech and Learning to Read, Elkridge, Md., May 1971. To appear in Language by Ear and by Eye: The Relationships between Speech and Reading, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press, in press).

---

*Most of the contents of this report, SR-27, are included in this listing.

Misreading: A Search for Causes. D. Shankweiler and I.Y. Liberman. Paper presented at the Conference on Communicating by Language: The Relationships between Speech and Learning to Read, Elkridge, Md., May 1971. To appear in Language by Ear and by Eye: The Relationships between Speech and Reading, J.F. Kavanagh and I.G. Mattingly, eds. (Cambridge, Mass.: M.I.T. Press, in press).

The Activity of the Adductor Laryngeal Muscles in Respect to Vowel Devoicing in Japanese. H. Hirose. Phonetica (1971) 23, 156-170.*

On the Evolution of Human Language. P. Lieberman. Plenary paper presented at the VIIth International Congress of Phonetic Sciences, Montreal, Canada, 25 August 1971. To appear in the Proceedings, in press.

Further Experimental Studies of Fundamental Frequency Contours. M. Studdert-Kennedy and K. Hadding. Paper presented at the VIIth International Congress of Phonetic Sciences, Montreal, Canada, August 1971. To appear in the Proceedings, in press.

Glottal Modes in Consonant Distinctions. L. Lisker and A.S. Abramson. Paper presented at the VIIth International Congress of Phonetic Sciences, Montreal, Canada, August 1971. To appear in the Proceedings, in press.

Voice Timing in Korean Stops. A.S. Abramson and L. Lisker. Paper presented at the VIIth International Congress of Phonetic Sciences, Montreal Canada, August 1971. To appear in the Proceedings, in press.

Dichotic Backward Masking of Complex Sounds. C.J. Darwin. Quarterly Journal of Experimental Psychology (November 1971) 23, Part 4, in press.

Distinctive Features and Laryngeal Control. L. Lisker and A.S. Abramson. Language (December 1971), in press.

Subsequent Recognition of Items Subject to Proactive Interference in Short-Term Memory. M.T. Turvey, D.L. Mosher, and L. Katz. Psychonomic Science, in press.

Preceding Vowel Duration as a Cue to the Perception of Word-Final Consonants. L.J. Raphael. Journal of the Acoustical Society of America, in press.

Current Reading Machine Research at Haskins Laboratories. J.H. Gaitenby. Stride (publication of the American Council of the Blind, Maryland) 2, in press.

Speech Cues and Sign Stimuli. I.G. Mattingly. American Scientist, in press.

---

*This paper reports research undertaken at the University of Tokyo. It is listed here because of its relevance to work being done at Haskins Laboratories and because the author is currently a visiting member of the staff.

Reports and Oral Papers

Backward Masking: Interruption or Integration? M. Turvey. University of Waterloo, April 1971. Also, Yale University, April 1971

Two Operations in Vision: Inferences from Masking by Noise and Patterns. M. Turvey. Aphasia Research Center, Boston VA Hospital, May 1971. Also, Ohio State University, May 1971.

Looking at the Larynx During Running Speech. Franklin S. Cooper. Oral paper presented at the 92nd Annual Meeting of the American Laryngological Association, San Francisco, Calif., 24 May 1971.

Two Operations in Visual Information Processing Prior to Short-Term Storage. M. Turvey. Lake Arrowhead Conference on Verbal Learning and Memory, Lake Arrowhead, Calif., 12-18 June 1971.

A Semi-Random Walk Through the Haskins Laboratories. R.S. Day. Bell Telephone Laboratories, 23 July 1971.

Language Codes and Memory Codes. A.M. Liberman, I.G. Mattingly, and M.T. Turvey. Paper presented at meeting on Coding Theory in Learning and Memory, Woods Hole, Mass., August 1971.

Cerebral Dominance for Speech in Relation to Handedness. D.P. Shankweiler. Aphasia Research Center, Boston VA Hospital, 24 September 1971.

The Concept of Developmental Dyslexia: Examination and Critique. D.P. Shankweiler. Paper presented at the Annual Meeting of the Academy of Aphasia, Denver, Colo. 18-19 October 1971.

An EMG Study of Japanese Accent Patterns. H. Hirose, Z. Simada, M. Sawashima, and O. Fujimura. Paper presented at the VIIth International Congress on Acoustics, Budapest, Hungary, 21 August 1971.*

Interactions Between Linguistic and Nonlinguistic Processing. R.S. Day and C.C. Wood. Paper presented at the 82nd meeting of the Acoustical Society of America, Denver, Colo., October 1971.

Perception of Linguistic and Nonlinguistic Dimensions of Dichotic Stimuli. R.S. Day, J.E. Cutting, and P.M. Copeland. Paper presented at the meeting of the Psychonomic Society, St. Louis, Mo., November 1971.

Thesis

On the Nature of Categorical Perception of Speech Sounds. David Bob Pisoni. Doctoral dissertation, The University of Michigan, Ann Arbor, 1971.

---

*This paper reports research undertaken at the University of Tokyo. It is listed here because of its relevance to work being done at the Laboratories and because the first author is currently a visiting member of the staff.